



Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Facultad de Farmacia y Bioquímica

Escuela Profesional de Ciencia de los Alimentos

**Secuenciamiento masivo (RNA-seq) y bioinformática
del transcriptoma de tejido graso de *Gallus gallus*
procedentes de centros de venta de Lima**

TESIS

**Para optar el Título Profesional de Licenciado en Ciencia y
Tecnología de los Alimentos**

AUTOR

Daniel Francisco VEGA OLCESE

ASESOR

Mg. Oscar ACOSTA CONCHUCOS

Lima, Perú

2019



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Vega D. Secuenciamiento masivo (RNA-seq) y bioinformática del transcriptoma de tejido graso de *Gallus gallus* procedentes de centros de venta de Lima [Tesis]. Lima: Universidad Nacional Mayor de San Marcos, Facultad de Farmacia y Bioquímica, Escuela Profesional de Ciencia de los Alimentos; 2019.

HOJA DE METADATOS COMPLEMENTARIOS

INFORMACIÓN GENERAL	
Título del Proyecto	"Secuenciamiento masivo (RNA-seq) y bioinformática del transcriptoma de tejido graso de <i>Gallus gallus</i> procedentes de centros de venta de Lima."
Área de investigación (*)	Ciencias Bioquímicas (Biotecnología)
Ubicación geográfica donde se desarrolla la investigación (incluir localidades y/o coordenadas geográficas)	Región Lima, Provincia Lima.
Institución que financia si corresponde	Facultad de Farmacia y Bioquímica - UNMSM
Año o rango de años que abarcó	2018 - 2019
DATOS DEL TESISISTA	
Apellidos y Nombres	Vega Olcese, Daniel Francisco
Número de matrícula	11040113
DNI	73067648
Indicar si es egresado o si aún está cursando estudios, de ser así especificar el año de estudios	Egresado
Código ORCID (opcional)	No
DATOS DEL ASESOR	
Apellidos y nombres	Oscar Acosta Conchucos
Código docente: 0A1914	Categoría: Asociado TC
Máximo grado alcanzado	Magister en Bioquímica
Código ORCID (obligatorio)	0000-0002-1912-0251
Título profesional	Biólogo con especialidad en Genética
Departamento Académico al que pertenece	Bioquímica (Facultad de Farmacia y Bioquímica)
Instituto de Investigación al que pertenece	Química Biológica, Microbiología y Biotecnología Marco Antonio Garrido Malo.
Grupo de investigación al que pertenece indicar si es coordinador, miembro o adherente del grupo de investigación	GENOBIDC (Coordinador)

(*) Según documentos oficiales de la Facultad



Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Farmacia y Bioquímica
Decanato



ACTA DE SUSTENTACIÓN DE TESIS

Los Miembros del Jurado Examinador y Calificador de la Tesis titulada:

Secuenciamiento masivo (RNA-seq) y bioinformática del transcriptoma de tejido graso de *Gallus gallus* procedentes de centros de venta de Lima"

Que presenta el Bachiller en Ciencia y Tecnología de los Alimentos:

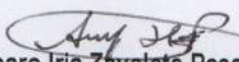
DANIEL FRANCISCO VEGA OLCESE


Que reunidos en la fecha se llevó a cabo la **SUSTENTACIÓN** de la **TESIS**, y después de las respuestas satisfactorias a las preguntas y objeciones formuladas por el Jurado, y practicada la votación han obtenido la siguiente calificación:


Quince (15) Bueno

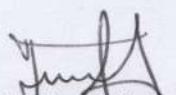
en conformidad con el Art. 34.º del Reglamento para la obtención del Grado Académico de Bachiller en Ciencia y Tecnología de los Alimentos y Título Profesional de Licenciado (a) en Ciencia y Tecnología de los Alimentos de la Facultad de Farmacia y Bioquímica de la Universidad Nacional Mayor de San Marcos.

Lima, 10 de mayo de 2019.


Dra. Amparo Iris Zavaleta Pesantes
Presidente


Dra. Mirtha Roque Alcarraz
Miembro


Mg. Carmen Gladys Peña Suasnabar
Miembro


Dra. Karim Lizeth Jiménez Aliaga
Miembro

"FARMACIA ES LA PROFESIÓN DEL MEDICAMENTO, DEL ALIMENTO Y DEL TÓXICO"



DEDICATORIA

*A Mamá Angela, Papá Daniel, y mi hermano Piero,
la mejor familia en el mundo, que con su apoyo, amor
y sacrificio fueron el soporte para poder culminar esta
linda etapa de mi vida con la presente investigación.*

AGRADECIMIENTOS

A la Escuela Profesional de los Alimentos de la prestigiosa Universidad Nacional Mayor de San Marcos, por permitirme llevar en el recuerdo vivencias inolvidables con las excelentes personas que conocí y por los docentes que forman parte de ella, que gracias a su sapiencia y conocimientos, me han hecho un profesional íntegro.

A mi asesor de tesis, el gran maestro Oscar Acosta, hombre de mil batallas por lograr que la ciencia avance en nuestro país para que las venideras generaciones sean las que dirijan a nuestro país hacia la era moderna global y finalmente, por brindarme todo lo necesario para poder llevar a cabo este trabajo. ¡Eternamente agradecido con usted, profe!

A mi madre por su amor, cuidados y apoyo incondicional, a mi padre por sus sabias enseñanza y ejemplo de persona íntegra, y mi hermano por su sapiencia para darme consejos y apoyo en toda travesía propuesta en mi vida. Somos una verdadera familia que se ama como ninguna, les agradezco infinitamente por todo.

A Melissa, quien me motivó a empezar con esta investigación y no claudicar a pesar de los intentos fallidos, por demostrarme que el verdadero amor existe en base a confianza y apoyo mutuo, y que los sueños, con perseverancia, se hacen posibles siempre.

Al Doctor Francisco Peñagaricano y el equipo del laboratorio de Genómica Cuantitativa Animal y Biología Computacional en el Departamento de Ciencias Animales en la Universidad de Florida, por mostrarme el camino de las nuevas técnicas y herramientas bioinformáticas para el análisis de las ómicas. Estoy seguro que seguirán encontrando respuestas a tan complejas preguntas biológicas.

A Dios, que mediante esa fuerza divina inexplicable, te alienta a seguir a pesar de las adversidades y dificultades de la vida.

Al Vicerrectorado Académico de Investigación (VRI) de la Universidad Nacional Mayor de San Marcos, por el financiamiento para el desarrollo de este trabajo, con el apoyo del Fondo de promoción de trabajo de tesis de pregrado (Código: A17040834a).

ÍNDICE GENERAL

	Páginas
DEDICATORIA	1
AGRADECIMIENTOS	2
ÍNDICE GENERAL	3
LISTA DE ABREVIATURAS	7
LISTA DE FIGURAS	8
LISTA DE TABLAS	9
RESUMEN	10
ABSTRACT	11
I. INTRODUCCIÓN	12
II. MARCO TEÓRICO	16
2.1 Situación general de los Pollos de Engorde en el Perú	16
2.1.1 Líneas Comerciales	18
2.1.2 Clasificación de los Pollos de Engorde	19
2.1.3 Subsectores del Pollo de Engorde en el Perú y Sudamérica	19
2.1.4 Empresas productoras	20
2.1.5 Canales de distribución	21
2.1.6 Factores determinantes del estado de la Industria Avícola	24
2.2 <i>Gallus gallus</i>	25
2.2.1 Genoma de <i>Gallus gallus</i>	26
2.2.2 Tejido adiposo en <i>Gallus gallus</i>	28

2.2.3 Expresión génica en el tejido adiposo abdominal y relación con el metabolismo	28
2.3 Ómicas	30
2.3.1 Importancia de la aplicación de ómicas en el análisis de alimentos	31
2.4 Transcriptómica	32
2.4.1 Tecnología para el análisis del transcriptoma	33
2.4.1.1 Secuenciamiento del ARN (RNA sequencing)	34
2.4.1.2 Tecnología aplicada para RNA-Seq: Secuenciamiento de Próxima Generación (NGS)	36
2.4.1.3 Tecnología de secuenciamiento Illumina/Solexa	40
2.4.2 Análisis Bioinformático	40
2.4.2.1 Control de calidad	40
2.4.2.2 Alineamiento de lecturas	41
2.4.2.3 Cuantificación de Transcritos	42
2.4.2.4 Expresión Diferencial Génica	44
III. MATERIALES Y MÉTODOS	46
3.1 Lugar de realización	46
3.2 Muestra	46
3.2.1 Toma de muestra	47
3.2. Análisis molecular	47
3.2.1 Tratamiento de la muestra	47
3.2.2 Extracción de ARN total	48

3.2.3 Secuenciamiento del Transcriptoma (Protocolo Illumina)	48
3.3 Análisis Bioinformático	49
3.3.1 Control de calidad de las lecturas secuenciadas	51
3.3.2 Alineamiento de las lecturas: TopHat	52
3.3.3 Ensamblaje de las lecturas: Cufflinks	54
3.3.4 Análisis de la expresión génica diferencial: Cuffdiff	55
3.3.5 Análisis de rutas las metabólicas y funciones fisiológicas de los genes diferencialmente expresados (upregulated y downregulated): KEGG	58
IV. RESULTADOS	59
4.1 Toma de muestra	59
4.2 Extracción de ARN	59
4.3 Análisis Bioinformático	61
4.3.1 <i>Raw data</i>	62
4.3.2 Control de calidad de las lecturas	63
4.3.3 Alineamiento de lecturas: Tophat	64
4.3.4 Ensamblaje de lecturas: Cufflinks	69
4.3.5 Expresión Diferencial Génica (DEG) en Cuffdiff	73
V. DISCUSIÓN	81
VI. CONCLUSIONES	94
VII. RECOMENDACIONES	95

VIII.	REFERENCIAS BIBLIOGRÁFICAS	96
IX.	ANEXOS	114

LISTA DE ABREVIATURAS

ADN: Ácido desoxirribonucleico

ADNc: Ácido desoxirribonucleico complementario

ARN: Ácido ribonucleico

ARNm: Ácido ribonucleico mensajero

ARNnc: Ácido ribonucleico no codificante

APA: Asociación Peruana Avicultura

BAM: Binary Alignment Map (mapa del alineamiento binario)

DEG: Differentially Expressed Genes (genes diferencialmente expresados)

FPKM: Fragments Per Kilobase Million (fragmentos por millón de kilobases)

IUM: Initially Unmapped reads (lecturas inicialmente no mapeadas)

GTF: Gene transfer format (formato de genes transformados)

KEGG: Kyoto Encyclopedia of Genes and Genomes (Enciclopedia de genes y genomas de Kioto)

MINAGRI: Ministerio de Agricultura y Riego

NGS: Next Generation Sequencing (Secuenciamiento de Próxima Generación)

PE: Paired End (final pareado)

RNA-seq: RNA sequencing

RPKM: Reads Per Kilobase Million (lecturas por millón de kilobases)

SAM: Sequence Alignment Map (Mapa del alineamiento de la secuencia)

SENASA: Servicio Nacional de Sanidad Agraria

SE: Single End (final singular)

TLR: Receptores de Tipo Toll

TPM: Transcripts Per Kilobase Million (transcritos por millón de kilobases)

LISTA DE FIGURAS

Figura 1. Relación de análisis de las ómicas. Tomada de Hasin <i>et al.</i> , 2017	31
Figura 2. Análisis de la Alimentómica del alimento desde la producción hasta el consumo. Modificado de Andjelkovic <i>et al.</i> , 2017	32
Figura 3. Experimento típico de RNA-Seq. Tomado de Zhong <i>et al.</i> , 2009	36
Figura 4. Mecanismos de acción de lo aplicado por los software TopHat y Cufflinks. Tomada de Trapnell <i>et al.</i> , 2010	43
Figura 5. Ejemplo de lectura obtenida de formato FastQ. Tomada de Bacci, 2015	50
Figura 6. Gráficos del Control de calidad de la <i>Raw Data</i> , generados por el programa FastQC. Tomado de Babraham bioinformatics, 2010	52
Figura 7. Esquema de la información que cada software del protocolo “Tuxedo” (tradicional) se encarga de procesar bajo dos condiciones que pueden ser control y experimental. Modificado de Trapnell <i>et al.</i> , 2012	57
Figura 8. Gel de agarosa con banda de ARN de muestra CA001, visualizada en el transiluminador LED luz azul.	58
Figura 9. Gráficos del control de calidad de los trasncritos secuenciados, generados por el programa FastQC	63
Figura 10. Deleciones en TopHat de muestra CA001	67
Figura 11. Inserciones en TopHat de muestra CA001	67
Figura 12. Uniones de empalme en TopHat (Muestra CA001)	68
Figura 13. Uniones Aceptadas en TopHat de muestra CA001	68
Figura 14. Expresión Génica en Cufflinks de Muestra CA001	70
Figura 15. Expresión de Transcritos en Cufflinks de Muestra CA001	71
Figura 16. Transcritos ensamblados en Cufflinks de Muestra CA001	72

LISTA DE TABLAS

Tabla N°1. Ranking de Producción de Carne de Pollo en América en 2013 (Miles de Toneladas).	20
Tabla N°2. Directorio de Centros de Acopio Y Comercialización de Aves en Pie – Lima.	23
Tabla N°3. Factores para determinar grado de desarrollo de la industria de pollos de engorde por país.	24
Tabla N°4. Clasificación taxonómica de <i>Gallus gallus</i> .	26
Tabla N°5. Principales Plataformas de Secuenciamiento usadas en la actualidad.	39
Tabla N°6. Parámetros utilizados por el software Cuffdiff para el análisis diferencial génico.	56
Tabla N°7. Cantidad Total de ARN y valor de RIN de las muestras analizadas.	60
Tabla N°8. Características de la tecnología utilizada para el secuenciamiento de las muestras CA001 y MB001.	61
Tabla N°9. Resumen de la base de datos obtenida del secuenciamiento NGS.	62
Tabla N°10. Cantidad de genes expresados en muestras CA001 y MB001 por Cufflinks.	73
Tabla N°11. Evaluación de la expresión diferencial génica (p ajustado < 0.05) de CA001 con respecto a muestra MB001	73
Tabla N°12. Expresión génica diferencial (sub-regulados) más relevantes de la muestra CA001 con respecto a la muestra MB001 de tejido adiposo de <i>Gallus gallus</i>	74
Tabla N°13. Expresión génica diferencial (sobre-regulados) de la muestra CA001 con respecto a la muestra MB001 de tejido adiposo de <i>Gallus gallus</i>	77
Tabla N°14. Expresión génica diferencial más relevantes (con expresión igual a cero, $\text{Log}_2 = \text{inf}$), de la muestra CA001 con respecto a la muestra MB001 de tejido adiposo de <i>Gallus gallus</i>	78
Tabla N°15. Expresión diferencial, no significativa, de genes representativos (lipogénicos e inmunológicos) en muestra CA001 con respecto a muestra MB001	80

RESUMEN

La carne de pollo es la fuente de proteína animal de mayor consumo por el ciudadano peruano. El consumo de pollo, solamente en la provincia de Lima alcanza los 58 kilos per cápita por año, y a nivel nacional, el promedio anual alcanza los 28 kilos per cápita, confirmando la preferencia de este producto. Bajo esta perspectiva, en los últimos años se han desarrollado técnicas biotecnológicas para optimizar los rendimientos y lograr un producto inocuo y con el valor nutricional correspondiente, es así que se han venido aplicando las tecnologías ómicas, en particular para el análisis e interpretación de los genomas y de expresión génica (transcriptómica), considerando factores externos a los que está sometido el animal como suplementos en la dieta o temperatura del ambiente donde crece. El objetivo de la investigación fue evaluar mediante secuenciación masiva (RNA-seq) y análisis bioinformático, las variaciones en el transcriptoma de tejido graso en estado fresco de *Gallus gallus* procedentes de centros de venta de Lima. La muestra CA001 proviene del Centro de Acopio de San Luis y la muestra MB001 del Mercado Modelo de Bellavista. Se secuenciaron las muestras de ARN total obtenidas del tejido adiposo de *Gallus gallus* con la tecnología Illumina NovaSeq 6000. Posteriormente, se realizó el análisis bioinformático utilizando el flujo de trabajo de una plataforma disponible en web y además con análisis complementarios utilizando herramientas bioinformáticas adaptadas al laboratorio. Para el mapeo de las lecturas se utilizó el software TopHat, luego el ensamblado con el software Cufflinks, y después el análisis por Cuffdiff, que utilizó el estadístico método de dispersión ciega (Blind Dispersion Method) para así obtener la expresión génica diferencial de la muestra CA001 con respecto a la muestra MB001. De un total de 74882 genes expresados por Cufflinks para MB001 y 91993 genes para CA001, se realizó el análisis de expresión diferencial encontrándose 57 genes sub-regulados, 6 sobre-regulados y 71 genes activos en CA001 con respecto a MB001, todos con diferencias significativas (p ajustado < 0.05) principalmente de rutas metabólicas de ácidos grasos, proteínas, resistencia a enfermedades y desarrollo celular. En general, se evidencian diferencias en el transcriptoma de ambas muestras de *Gallus gallus*, posiblemente asociados a suplementos dietarios u otros factores implicados en la crianza de este recurso avícola de elevado consumo y de gran impacto en la industria alimentaria.

Palabras clave: *Gallus gallus*, transcriptómica, RNA-seq, mapeo, alineamientos, expresión diferencial génica.

ABSTRACT

Chicken meat is the source of animal protein most consumed by the Peruvian population. The consumption of chicken, only in the province of Lima reaches 58 kilos per capita per year, and nationally, the annual average reaches 28 kilos per capita, confirming the preference of this product. Under this perspective, biotechnological techniques have been developed in recent years to optimize yields and achieve a safe product with the corresponding nutritional value. This is the reason why omic technologies have been applied, particularly for the analysis and interpretation of genomes and gene expression (transcriptomics), considering external factors to which the animal is subjected as supplements in the diet or temperature of the environment where it grows. The objective of the research is to evaluate, through massive sequencing (RNA-seq) and bioinformatic analysis, the variations in the transcriptome of fresh fatty tissue of *Gallus gallus* from Lima sales centers. The sample CA001 comes from San Luis Collection Center and MB001 from Model Market of Bellavista. The total RNA samples obtained from the *Gallus gallus* adipose tissue were sequenced with the Illumina NovaSeq 6000 technology. Subsequently, the bioinformatic analysis was performed using the workflow of a platform available on the web and also with complementary analyzes using bioinformatics tools adapted to the laboratory. For the mapping of the readings, the software TopHat was used, then the assembly with the software Cufflinks, and then the analysis by Cuffdiff, which used the statistical method of blind dispersion in order to obtain the differential gene expression of the sample CA001 with respect to sample MB001. From a total of 74882 genes expressed by Cufflinks for MB001 and 91993 genes for CA001, the differential expression analysis was performed, finding 57 downregulated genes, 6 upregulated genes and 71 active genes in CA001 with respect to MB001, all with significant differences (p adjusted <0.05) mainly of metabolic pathways of fatty acids, proteins, disease resistance and cell development. In general, there are differences in the transcriptome of both *Gallus gallus* samples, possibly associated with dietary supplements or other factors involved in the raising of this high consumption and high impact poultry resource in the food industry.

Key words: *Gallus gallus*, transcriptomics, RNA-seq, mapping, alignments, gene differential expression.

I. INTRODUCCIÓN

El consumo de pollo, solamente en la provincia de Lima alcanza los 58 kilos per cápita por año, y a nivel nacional, el promedio anual alcanza los 28 kilos per cápita, según datos brindados por la Asociación Peruana de Avicultura (1). A fines de noviembre de 2018, la industria avícola participaba con el 25% dentro del Valor Bruto del Sector Agropecuario y perdura como el principal proveedor de proteína animal en el Perú. La producción avícola para noviembre del 2018, siendo el último reporte del MINAGRI hasta la finalización de esta tesis, mostró un crecimiento del 6.9% en relación al mismo mes en el 2017. Este crecimiento se basó sustancialmente en la producción de carne de pollo y huevo de gallina. Haciendo una comparación con el periodo de enero a noviembre del 2017, la producción avícola, en el mismo periodo del 2018, creció un 7.7%, siendo las principales regiones productoras de carne de pollo: Lima (54.1%), La Libertad (18.4%), Arequipa (10.0%) e Ica (4.6%) respectivamente (2). Esto confirma la premisa de que el pollo es la fuente de proteína animal más consumida y de mayor preferencia por el consumidor peruano.

Debido a la importancia de esta actividad y al consumo masivo de este alimento, es necesario asegurar la calidad nutricional y las buenas prácticas de crianza en la su proceso de crecimiento, lo cual es complicado evaluar in situ. Existen varias herramientas o procesos que permiten verificar la calidad, entre ellas el análisis genético. Con el avance de la biotecnología moderna se han desarrollado técnicas para evaluar la calidad entre las que destacan la genética molecular y las ómicas, que, siendo esta última de gran impacto, actualmente viene siendo utilizada en el ámbito del desarrollo e investigación de diferentes especies de animales que serán destinados para el consumo humano.

Los resultados obtenidos por la secuenciación del genoma animal han conllevado el desarrollo de una plétora de técnicas analíticas para analizar la variación total de la expresión de genes, proteínas y metabolitos. Se conoce, que la información genética se mantiene constante durante la vida de un animal, lo que cambia, son los niveles de expresión de transcritos (transcriptómica), proteínas (proteómica) y

metabolitos (metabolómica); generalmente, estos cambios están regulados y supeditados a una diversidad de factores ambientales y fisiológicos (3).

La transcriptómica, es una de las principales tecnologías ómicas que se utiliza actualmente para investigar el perfil de expresión de genes. Su objeto de estudio, el transcriptoma, comprende a todos los transcritos de ARNm provenientes de la transcripción del ADN. Esto representa ese pequeño porcentaje de genes que son transcritos en moléculas de ARN, lo que se ha estimado, por ejemplo, que es menor al 5% en el genoma humano (4). La finalidad de la transcriptómica es identificar los niveles de los genes en las transcripciones del ARNm en respuesta a diferentes factores externos como la temperatura y los nutrientes propios de la dieta, o condiciones fisiopatológicas específicas, así como identificar los genes que pertenecen a rasgos específicos. Es, hasta el momento, una de las herramientas más importantes para poder entender y encontrar respuestas moleculares a la adición en la dieta de diferentes suplementos alimenticios, a enfermedades, como la miopatía muscular, que afectan la calidad final del producto y otros aspectos relacionados a la inmunidad y resistencia a enfermedades (3).

Los avances en la tecnología de secuenciación de próxima generación (NGS) han aumentado sustancialmente el número de especies, como el pollo (*Gallus gallus*), con ensambles completos de genoma de alta calidad. Estos avances también han permitido estudiar la funcionalidad del transcriptoma mediante su secuenciamiento masivo denominado RNA-seq (5).

El contenido de grasa intramuscular en la carne de pollo es un elemento sumamente relevante para las preferencias del consumidor debido a su correlación positiva con la calidad de la carne, la que incluye jugosidad y frescura (6). La diferenciación adipocítica y el metabolismo de los lípidos juegan un rol significativo en el proceso de deposición de la grasa intramuscular (IMF), lo que demanda una sucesión de cambios en la regulación de diferentes factores de transcripción y la regulación de los genes que controlan el metabolismo de los lípidos (7). La actividad del factor de transcripción está regulada por una variedad de factores intracelulares y extracelulares y de vías de señalización, que desempeñan un papel determinante en la diferenciación de los adipocitos

precursores en adipocitos maduros (8). La generación de grasa de aves de corral es diferente de la generación de grasa de mamíferos, debido a que los ácidos grasos de *Gallus gallus* se sintetizan principalmente en el hígado. Las lipoproteínas hidrolizan y liberan ácidos grasos que se han transportado al tejido adiposo, luego se sintetizan y se depositan en las células adiposas (9). La investigación sobre los mecanismos de regulación molecular de la generación de grasa es clave para comprender los mecanismos moleculares de la calidad de la carne.

Toda esta información aporta sobre la importancia de estudiar no solo genes sino los niveles de expresión génica en todo el genoma (transcriptoma). La elección de las muestras es un aspecto relevante, siendo apropiado el uso de muestras de tejido hepático, graso y muscular para el análisis de expresión debido a su importancia biológica y económica (10).

Por estas razones y para ampliar el conocimiento sobre el tema en nuestro país, se plantea realizar el estudio de la expresión diferencial génica de ARN extraído de tejido graso de *Gallus gallus* (pollos de engorde) procedentes de diferentes centros de venta de Lima, considerando diferentes condiciones de inocuidad y calidad, para realizar el secuenciamiento de última generación (NGS) y el análisis bioinformático del transcriptoma en el contexto de las ómicas.

HIPÓTESIS

- Existen diferencias en el transcriptoma de tejido graso de *Gallus gallus* de centros de venta de Lima.

OBJETIVO GENERAL

- Evaluar mediante secuenciamiento masivo (RNA-seq) y análisis bioinformático, las variaciones en el transcriptoma de tejido graso de *Gallus gallus* procedentes de centros de venta de Lima.

OBJETIVOS ESPECÍFICOS

- Realizar el secuenciamiento masivo (RNA-seq) de muestras de tejido graso de *Gallus gallus*
- Realizar el análisis bioinformático del transcriptoma obtenido por secuenciamiento de tejido graso de *Gallus gallus*.
- Establecer las diferencias o variaciones entre los niveles de expresión génica de tejido graso de *Gallus gallus* según lugar de procedencia.

II. MARCO TEÓRICO

2.1 Situación general de los Pollos de Engorde en el Perú

El año 2016 el PBI del país creció en un 3.9%, catapultado por actividades como la minería e hidrocarburos (16.3%), electricidad y agua (7.3%), servicios (3.9%), agropecuario (2.0%), entre otros (11).

A raíz de efectos climatológicos que afectaron al recurso hídrico, impactando en cultivos como el trigo, maíz, entre otros, el sector agropecuario se vio debilitado respecto al 3.2% alcanzado en 2015 (12).

En el año 2016, apoyado por el crecimiento en los volúmenes de producción de porcinos que fue 7.21% (5.34% acumulado), huevos 6.08% (3.81% acumulado), aves 5.57% (5.22% acumulado), entre otros, el sector pecuario colaboró con el 3.9% (en contra del 0.7% del sector agrícola). La mayor producción de Pollo de Engorde (4.30%), pavo (3.00%), entre otros, principalmente, produjo un aumento en la producción de aves. Sumado a ello, para diciembre de 2016, el volumen de las importaciones de carne de ave se ubicó en 4.81 miles t, 12.4% menor a diciembre de 2015, así como también, el volumen de las exportaciones llegó a las 178 t, lo que representó un incremento de 272.4% (12).

La producción de aves creció un considerable 5.22% (alcanzando 1, 827 miles t), por la mayor producción de pollitos bebé, principalmente en ciudades como Lima, con un 9.35%, Arequipa 4.15%, La Libertad 0.13%, entre otros. Es importante recalcar que las ciudades antes mencionadas, representan más del 80% de la producción de aves en el Perú. Con lo que respecta al precio, este también se vio afectado; el precio mayorista del Pollo de Engorde a diciembre de 2016 llegó a los S/ 5.20 / kg en Lima y Callao, dándose así, un aumento de 1.90% en relación a lo que costaba en diciembre de 2015 (13).

Para noviembre de 2018 se indicó que la industria avícola nacional viene participando con el 25% dentro del Valor Bruto del Sector Agropecuario, siendo el más importante proveedor de proteína animal de nuestro país (2).

La producción de pollos de engorde representa un segmento muy significativo en la industria de animales que son fuente de carne para el consumo humano, debido a que los pollos de engorde pueden ser el productor más eficiente de carne magra en comparación con las especies de cerdos y bovinos (14,15).

La producción en el mes de noviembre del año 2018, mostró un crecimiento del 6.9% con relación al mismo mes del año anterior. Esto se debió a la producción principalmente de carne de pollo y huevo de gallina. En esa relación de meses (enero a noviembre) del 2018 comparado con la misma del año anterior, la producción avícola incrementó el 7.7%. En noviembre, las importaciones de carne de ave (entre ellos pollo, gallina, pavo) llegaron a un total de 6602 toneladas, y representa un 27% más de lo importado en el mismo mes del 2017 (2).

El más significativo proveedor de proteína animal en nuestro país es el sector avícola, de acuerdo a los que señala José Vera, presidente de la APA (Asociación Peruana de Avicultura), el 67% de la población peruana consume este tipo de productos. El consumo de carne de pollo de engorde en el Perú posee una tendencia exponencial, debido a que se ha elevado en un promedio de 20 kg por persona al año a comienzos de los primeros meses del 2000 a más de 40 kg a 2016, llegando a darse valores máximos de 70 kg por persona, principalmente en la capital, Lima (16).

Rolando Arellano, presidente de Arellano Marketing, una de las principales empresas dedicadas a la investigación de preferencias del consumidor y mercados de consumo, en su artículo “El Pollo de Cada Día”, remarcó que el Perú en Latinoamérica, es el segundo consumidor de pollo per cápita, superado solamente por Brasil. Allá por la década de los setenta es que se da inicio a este consumo masivo de pollo. Algunos pioneros comenzaron a criar este animal de forma estructurada, ordenada y con un tinte científico bastante importante, consiguiendo que para la mayoría de la población su adquisición sea de fácil acceso (17).

Son 450000 cargos, los que generan en el Perú las empresas (cerca de 52) del subsector de pollo de engorde, que consiguen niveles de ventas de hasta la cantidad de S/.8100 millones y márgenes de alrededor del 3.00% (18).

2.1.1 Líneas Comerciales

Los pollos utilizados para consumo humano, son denominados pollos de engorde (“broilers” en inglés), definidos como la variedad de pollos criados concretamente para la producción de carne. Estos pollos son alimentados específicamente para lograr que a largo de su desarrollo tengan una buena salud y un tamaño promedio para los estándares de calidad requeridos por el consumidor (19).

En el Perú destacan principalmente dos líneas genéticas de Pollo de Engorde como son Cobb y Ross (20). El 54% de las granjas están localizadas principalmente en la región Lima y Callao debido a las favorables condiciones climatológicas y a la accesibilidad del territorio; la región de La Libertad concentra el 19% y la región Arequipa el 10%, sumando entre las tres regiones el 83% del total de granjas en el Perú. Sin embargo, la costa concentra el 90% de la producción nacional de Pollos de Engorde (21). La producción de Pollos de Engorde en el Perú de enero a abril de 2017 fue de 558.2 millones aves, 5.8% más que lo producido en 2016 para el mismo periodo (22).

Los pollos de engorde o parrilleros son el resultado del cruce de gallos reproductores pesados con gallinas de doble propósito (producción de carne y huevo), entre ellas se encuentran (23):

a) Cobb Vantress

Son de crecimiento rápido, baja conversión alimenticia, alta viabilidad, alta rusticidad en el manejo y de factible adaptabilidad a cambios climáticos. Es la línea más explotada en la actualidad.

b) Ross Breeders

Son pollos de menos velocidad de crecimiento que la Cobb Vantress, pero con la conversión alimenticia menor. También conserva una alta rusticidad y adaptabilidad a diferentes climas.

c) Arbor Acres

En años anteriores fue la más popular, por los requerimientos del mercado (pollo grande) pero que después a causa de los cambios en la demanda de pollo más

pequeño (parrilleros) ha ido perdiendo mercado. Poseen baja conversión alimenticia, son rápidos de crecimiento, pero son exigentes y para adaptarse a condiciones climáticas exigen medidas de manejo costosas.

2.1.2 Clasificación de los Pollos de Engorde

Dentro de esta clasificación, los más comercializados y conocidos son los siguientes (24):

- **Pollo vivo o en pie:** es aquel que se produce durante 42 días aproximadamente en galpones de engorde y que se vende vivo a los mayoristas y/o centros de abasto en el mercado nacional.
- **Pollo procesado y derivados:** es el que luego del proceso de producción de 42 días, es beneficiado en las plantas que poseen las empresas avícolas como Redondos y San Fernando; posteriormente es envasado y vendido en el mercado con marca propia (el pollo que se expone en las cámaras de frío de los supermercados Metro y Tottus). Con respecto a los derivados, estos son aquellos productos que se generan a partir de la carne de pollo (jamonada de pollo, hamburguesa de pollo, etc.).
- **Pollo parrillero:** es aquel que se produce en menos de 40 días y que se dirige para comercialización en restaurantes, pollerías y lugares especializados en esta clase de platos en el país.

Solo los pollos procesados y parrillero, los que se comercializan y distribuyen en el canal moderno; es decir, en los restaurantes y supermercados, siendo aproximadamente el 20% de la producción y facturación de la industria, mientras el 80% restante se efectúa por el canal tradicional, el cual tiene un gran poder de negociación con el productor.

2.1.3 Subsectores del Pollo de Engorde en el Perú y Sudamérica

Existen en la actualidad, en Sudamérica, dos mercados diferenciados en lo que concierne al Subsector del Pollo de Engorde, estos son: (a) la industria productiva con valor agregado y (b) la industria productiva sin valor o bajo valor agregado. Entre las grandes industrias con valor agregado se encuentran las de Chile y Brasil como los máximos exponentes en lo que respecta a sanidad, calidad,

costos, índices productivos, cadena de distribución y productos finales con alto valor agregado. Sin embargo, Perú junto con Colombia y Ecuador, teniendo en cuenta su proyección de mercado y volumen de producción, son los países que se hallan en pleno desarrollo de la cadena productiva y también de la generación de valor agregado para poder comenzar con exportaciones a mercados dominantes del globo como el continente europeo o China (18).

Se calcula que, en el 2016, América tuvo una elevada producción siendo el 44% del Pollo de Engorde producido en el ámbito mundial, siendo el líder en Sudamérica, Brasil, con una participación del 29.5%, por debajo del EE.UU. con un 41.7%. Con esto, Brasil se mantiene al frente como el mercado con mayor desarrollo en el Subsector del Pollo de Engorde durante los últimos 20 años (25).

Tabla N°1. Ranking de Producción de Carne de Pollo en América en 2013 (Miles de Toneladas)

País	Total	Participación (%)
EE.UU.	17546	41.7
Brasil	12435	29.5
México	2801	6.7
Argentina	1779	4.2
Colombia	1276	3.0
Venezuela	1273	3.0
Perú	1203	2.9

Fuente: Modificado de Bustamante *et al.*, 2018 (18).

2.1.4 Empresas productoras

El sector avícola concentra el 58.25% de sus ventas en cuatro empresas, siendo la primera en el mercado San Fernando, seguido por Redondos, Ganadera Santa Elena y Chimú Agropecuaria. Este escenario genera la competitividad de la industria, haciendo que se busque constantemente hacer productos más competitivos, de calidad y con un valor agregado tecnológico y científico; incentivando en una inversión regular en: (a) tecnología para el mejoramiento del sistema de producción, (b) mano de obra, (c) capacitación de personal, e (d) investigación y desarrollo (26).

Un hecho importante relacionado a la competitividad en esta industria, fue la fusión que se realizó entre Ganadera Santa Elena S.A. (Familia Valdez Bernos) y Avinka S.A. (Continental Grain Company). Con esta operación la nueva compañía se convertiría en el número uno en utilidades y la segunda en ventas (diario Gestión 12/2016), aprovechando la experiencia y tecnología de Ganadera Santa Elena y la experiencia de Avinka en el beneficio y el procesamiento de la carne de Pollo de Engorde para su comercialización (27).

2.1.5 Canales de distribución

La distribución de pollo de engorde en el Perú presenta una elevada concentración en la costa, siendo Lima el departamento más sobresaliente. De los motivos principales que posee esta región son las adecuadas condiciones, llegando a presentar hasta un 93% de la producción nacional (21). Asimismo, si bien en el mercado peruano se tiene una clara preferencia por el “pollo fresco”, el cual los productores venden vivo a los centros de acopio, esto impide que el producto acceda en las condiciones necesarias hacia zonas muy alejadas en la sierra y la selva (28).

Los centros de acopio son los que comercializan el mayor porcentaje de pollos de engorde con un aproximado de 85%, el cual comercializan a los principales mercados minoristas, mientras el restante 15% se vende beneficiado y tiene como adquirentes a los autoservicios y restaurantes (29).

Se encuentra una acentuada orientación en la conformación de la cadena de distribución hacia el canal tradicional a través de los centros de acopio, lo que genera graves desventajas para la industria, tales como: el inapropiado control de la sanidad, al sustentarse de mataderos informales que terminan desechando casi la totalidad del valor sumado por el productor en la etapa de la crianza; y un marcado poder de negociación de los comerciantes mayoristas, lo que produce presión sobre los costos y el margen del productor, a través de la especulación, aprovechándose de la dificultad de este para almacenar el pollo vivo. Esta preferencia por el pollo fresco y la relevancia del canal tradicional supone para algunos productores una barrera para aumentar sus márgenes (18).

Tabla N°2. Directorio de Centros de Acopio Y Comercialización de Aves en Pie - Lima

Nombre y/o Razón Social	Dirección	Distrito	No. De Comerciantes
Asociación de Comerciantes Mayoristas de Aves Vivas La Victoria y Afines (ACOMAC)	Pasaje Mogrovejo No. 1690	La Victoria	12
Asociación de Comerciantes Mayoristas de Aves Vivas Acho-Rímac	Altura Puente Huánuco – Campo Ferial Acho	Rímac	14
Inmobiliaria de Comerciantes de Aves Vivas S.A.	Cruce Av. Industrial con Separadora	San Luis	12
San Fernando S.A.	Calle Las Fraguas	Independencia	20
Sin Razón Social	Calle Maranga. Cdra. 7 y Mercado Bellavista (Altura del hospital D.A. Carrión)	Callao	4
Sin Razón Social	Av. Universitaria No. 2895 y Mercado La Cumbre	Comas	6
Sin Razón Social	Pantanos de Villa	Chorrillos	3
Sin Razón Social	Mercado Unicachi	Urb. Pro-Los Olivos	4
Sin Razón Social	Mercado de Huamanga	Puente Piedra	5
Sin Razón Social	Jr. Inca Cdra. 6	Rímac	6
Sin Razón Social	Av. El Sol No. 309	Canto Grande – SJL	5
Sin Razón Social	Av. Huarochirí Cdra. 5	Santa Anita	4
Sin Razón Social	Av. La Costanera Cdra. 12	San Miguel	3
Sin Razón Social	Calle Adelcosur Mz. F, Lote 5 y 6. Urb. Parque Villa El Salvador	Villa El Salvador	8

Fuente: Estadística de Comercialización de Aves Vivas en Lima Metropolitana (30).

2.1.6 Factores determinantes del estado de la Industria Avícola

Es trascendental identificar cuáles son los factores determinantes que inciden en un producto o industria (naciente, desarrollo, maduración o declinante); esto para poder establecer e implementar las estrategias a seguir y propinar acciones que impacten de manera positiva el negocio en desarrollo (18).

En la siguiente tabla N°3, se podrán encontrar los factores que determinan la madurez de la industria del Perú con respecto a los otros actores inmersos en la misma dentro del continente, dándose a conocer cuáles son las falencias y en lo que hay que hacer hincapié en mejorar (18).

Tabla N°3. Factores para determinar grado de desarrollo de la industria de pollos de engorde por país.

Factor	Grado de madurez por país			
	Perú	Colombia	Brasil	Chile
Eficiencia en la producción	3	3	4	4
Mercados Internacionales	1	2	5	4
Sanidad e Inocuidad	2	2	4	4
Sostenibilidad	2	2	4	4
PROMEDIO	2	2.3	4.3	4

Nota. 1 – Inexistente; 2 – Bajo; 3 – Mediano; 4 – Alto; y 5 – Absoluto.

Fuente: Modificado de Bustamante *et al.*, 2018 (18).

2.2 *Gallus gallus*

El pollo (*Gallus gallus*), es una especie considerada como una interfase evolutiva entre los mamíferos y otros vertebrados y es útil como el principal modelo de laboratorio para un promedio de 9600 aves existentes. El pollo también representa al primer animal agrícola del que se ha alcanzado la secuencia de su genoma (31).

Los pollos se domesticaron en Asia, alrededor de 5400 años a.C., o quizás más temprano en el 8000 a.C. Darwin (32) refirió que las aves de la jungla roja eran el antepasado más cercano al pollo doméstico, una teoría que se confirmó posteriormente al realizarse el análisis del ADN mitocondrial (33).

La especie con mayor distribución a nivel mundial en la industria avícola, es el pollo doméstico (*Gallus gallus domesticus*), el cual provee al ser humano de una considerable fuente de proteína, incluyendo carne y huevo (34). Una gran cantidad de investigadores se han centrado en la reconstrucción de la historia matrilineal de los pollos domésticos utilizando datos de secuenciamiento de ADN mitocondrial (ADNmt). Dichos avances han permitido confirmar que el ave roja (red junglefowl/ *Gallus gallus*) es el ancestro silvestre del pollo doméstico (33,35).

Los pollos domésticos se clasifican en dos categorías principales; pollos comerciales y de pueblo indígena (36). En los países en desarrollo, estos últimos son fundamentales para el sustento de los pequeños agricultores, adaptándose a las condiciones externas locales. Sirven como aves de doble propósito (huevo y carne). En lo que concierne a líneas de pollo comerciales, los factores genéticos que controlan rasgos de desarrollo, crecimiento y reproducción han sido bien estudiados, mientras que el pollo indígena no ha tenido dicha (37,38).

Tabla N°4. Clasificación taxonómica de *Gallus gallus*

TAXONOMÍA DE GALLUS GALLUS	
Reino	Animalia
Subreino	Bilateria
Filum	Chordata
Subfilum	Vertebrata
Clase	Aves
Orden	Galliformes
Familia	Phasianidae
Género	Gallus
Especie	<i>Gallus gallus</i>

Fuente: modificado de <https://www.itis.gov/servlet/SingleRpt> (39)

2.2.1 Genoma de *Gallus gallus*

La secuencia del genoma de un organismo codifica tanto ARNnc y proteínas. El genoma del pollo proporciona nuevas perspectivas en la estructura y contenido de los genes de los mamíferos, así como el entendimiento del contenido del gen aviar y la evolución de los genes ARNnc (31).

El genoma del pollo es menos de la mitad del tamaño del ratón o humano, y está compuesto de 38 autosomas y dos cromosomas sexuales, denominados Z y W. A diferencia de los mamíferos, las aves macho son homocigotas (ZZ) mientras que las hembras son heterocigotas (ZW). Los cromosomas de pollo, al igual que otras especies aviares, presentan una variación mucho mayor en el rango de tamaño, que varía de 200 Mb a 5 Mb; Los cromosomas más pequeños se denominan "microcromosomas" (40).

La primera secuencia del genoma de *Gallus gallus* corresponde a la, parcialmente consanguínea, Gallina bankiva, siendo también el principal progenitor silvestre de los pollos domésticos y proveyendo un sustancial avance para la genética aviar (31). Con el pasar del tiempo. Se han realizado mayores aportes al ensamblaje del genoma de referencia aviar (41). Una segunda construcción (Gallus_gallus 2.1; GCA_000002315.1) del modelo original producido el 2004, mejorando la presentación del ensamblaje de las bases, el orden y la orientación, utilizando nuevos datos de mapeo de polimorfismo de un nucleótido (SNP) y la secuenciación dirigida de BACs y fásmidos. La secuencia total de Gallus_gallus 2.1 midió 1,09 Gb, aproximadamente el 95% de los cuales estaba anclada a los autosomas 1–28 y 32, junto con los cromosomas sexuales GGAZ y GGAW (7).

Una tercera construcción del genoma (Gallus_gallus-4.0; GCA_000002315.2), incluyendo el uso de la tecnología NGS (~12× de 454 secuencias de titanio) en combinación con las antiguas lecturas secuenciadas por el método de Sanger. Un incremento de tamaño de contigs N50 y scaffolds a 252 kb (460%) y 12.4 Mb, respectivamente. Además, el nuevo ensamble suprimió la mayor parte de duplicaciones artefactuales suscritas en evaluaciones comparativas anteriores (8), probablemente debido al hecho de que el ave secuenciada era solo parcialmente consanguínea y, por lo tanto, retenía regiones de diversidad alélica que fueron erróneamente ensamblados como loci independientes. La cantidad total de secuencia asignada a los cromosomas aumentó en 15 Mb, después de tener en cuenta los errores de duplicación (42).

El último aporte con respecto al genoma de *Gallus gallus* se da con el ensamble Gallus_gallus-5.0. Esto debido a que ensamblajes de genoma de alta calidad se han vuelto considerablemente necesarios para lograr obtener el mayor potencial de los estudios NGS, ya que los ensamblajes iniciales de "borrador" se han encontrado insuficientes para la visualización más completa de las contribuciones alélicas a rasgos particularmente complejos (42).

Comparado con Gallus_gallus-4.0, el ensamblaje de Gallus_gallus-5.0 añade un total de 183 Mb de secuencias nuevas, muestra un aumento de 10 veces en la longitud de contig N50 y agrega 4679 genes anotados, tanto transcritos de codificación de proteína como no transcritos (42).

2.2.2 Tejido adiposo en *Gallus gallus*

En los animales, incluyendo al ser humano, el blanco tejido adiposo almacena energía en forma de triglicéridos en exceso. Sería lo ideal reducir la deposición de tejido adiposo, ya que esto conllevaría a un aumento de la eficiencia de la alimentación y reduciría considerablemente los costos de producción. Sin embargo, es significativo considerar que el tejido adiposo secreta una variedad de proteínas que incluyen quimiocinas, leptina, adiponectina e interleucina IL – 6 e IL 8 (43).

La secreción de dichos factores a través del tejido adiposo genera un impacto relevante en diferentes procesos metabólicos y fisiológicos. Por ejemplo, la secreción de citoquinas y la expresión de receptores de tipo toll (TLR) regulan la respuesta del huésped a la infección e inmunidad (44). La tecnología de ADN complementario ha permitido el estudio de la expresión génica de factores generados por el tejido adiposo y otros factores relacionados a partir del enfoque de los microarrays, aunque aplicado de forma muy limitada para el análisis de los órganos en mención de pollos de engorde (45,46).

Otro tema que actualmente representa un reto de estudio, que está ligado con la especie en estudio, es la excesiva acumulación de tejido adiposo en gallinas ponedoras, lo que se ha convertido en un significativo problema en la producción de pollo, debido a que esta condición podría causar síndrome hemorrágico de hígado graso, teniendo como consecuencia el decrecimiento de la producción de huevos y la calidad nutricional de los mismos, en la que estaría inmiscuida la síntesis de ácidos grasos (47).

2.2.3 Expresión génica en el tejido adiposo abdominal y relación con el metabolismo

El tejido adiposo abdominal de *Gallus gallus* puede ser considerado un elemento inmunológico y análogo a la grasa visceral en otras especies. Además, la tasa de crecimiento extremo inducida genéticamente en pollos de engorde puede afectar la regulación positiva de los factores relacionados con el sistema inmunológico en la grasa abdominal (48).

En los últimos 44 años, la selección genética para obtener la delgadez ha aumentado la tasa de crecimiento en pollos de engorde 6 veces y se ha duplicado en el músculo de la pechuga (14,15). Posiblemente, este inesperado cambio en la tasa de crecimiento de los pollos de engorde haya producido una respuesta inflamatoria debido a que los procesos de crecimiento pueden haber sido estimulados más allá de los mecanismos homeostáticos regulares. Por lo tanto, la sobrerregulación de los factores relacionados con el sistema inmunológico puede haber sido provocada a causa de la inflamación inducida por las tasas de crecimiento extremas en los pollos de engorde. Aunque, una infección o agentes infecciosos sobrerregulan los factores relacionados con el sistema inmunológico, es poco probable que se involucre en función de la salud y la apariencia de los pollos (48).

Se ha considerado, según estudios en relación a la edad del pollo, que el tejido adiposo de *Gallus gallus* tendría estrecha relación con el sistema inmune y sería considerado un órgano inmunológico. Genes asociados a la ruta metabólica de TLRs (Receptor de Tipo Toll) y a la citotoxicidad mediada por células asesinas estarían sobre regulados en el tejido adiposo del pollo de engorde (48).

En cuanto a los genes lipogénicos como LPL (lipoproteinlipasa) y FAS (ácido graso sintasa), ven afectada la regulación de su expresión en relación a la adición de nutrientes, como la betaína, a la dieta control (47).

2.3 Ómicas

Esta palabra se refiere al campo de estudio, dentro de las ciencias biológicas, que al término de su nombre llevan el sufijo “ómica”, así como la genómica, transcriptómica, proteómica o metabolómica. Mientras el término “oma”, se agrega como sufijo del objeto de estudio de cada uno de estos campos, tal como el genoma, proteoma, transcriptoma o metaboloma, respectivamente (49).

En conclusión, el objetivo principal de las ciencias ómicas es identificar, caracterizar y cuantificar cada una de las moléculas biológicas que hacen la estructura, función y dinámica de una célula, tejido u organismo (49), como el transcriptoma proveniente del tejido adiposo de *Gallus gallus* en el caso de esta tesis.

Las tecnologías ómicas se caracterizan por la generación de gran cantidad de datos y el respectivo análisis bioinformático, como por ejemplo en el caso del estudio del transcriptoma, donde se evalúan los niveles diferenciales de expresión génica.

Los datos de ómicas se recopilan en todo el conjunto de moléculas, representadas como círculos. Excepto por el genoma, todas las capas de datos reflejan tanto la regulación genética como el ambiente en el que se desarrolla el individuo, lo que puede afectar a cada molécula individual en una medida diferente. Las flechas delgadas representan interacciones potenciales o correlaciones detectadas entre moléculas en diferentes capas; por ejemplo, la transcripción de la molécula roja se puede correlacionar con múltiples proteínas (Figura 1).

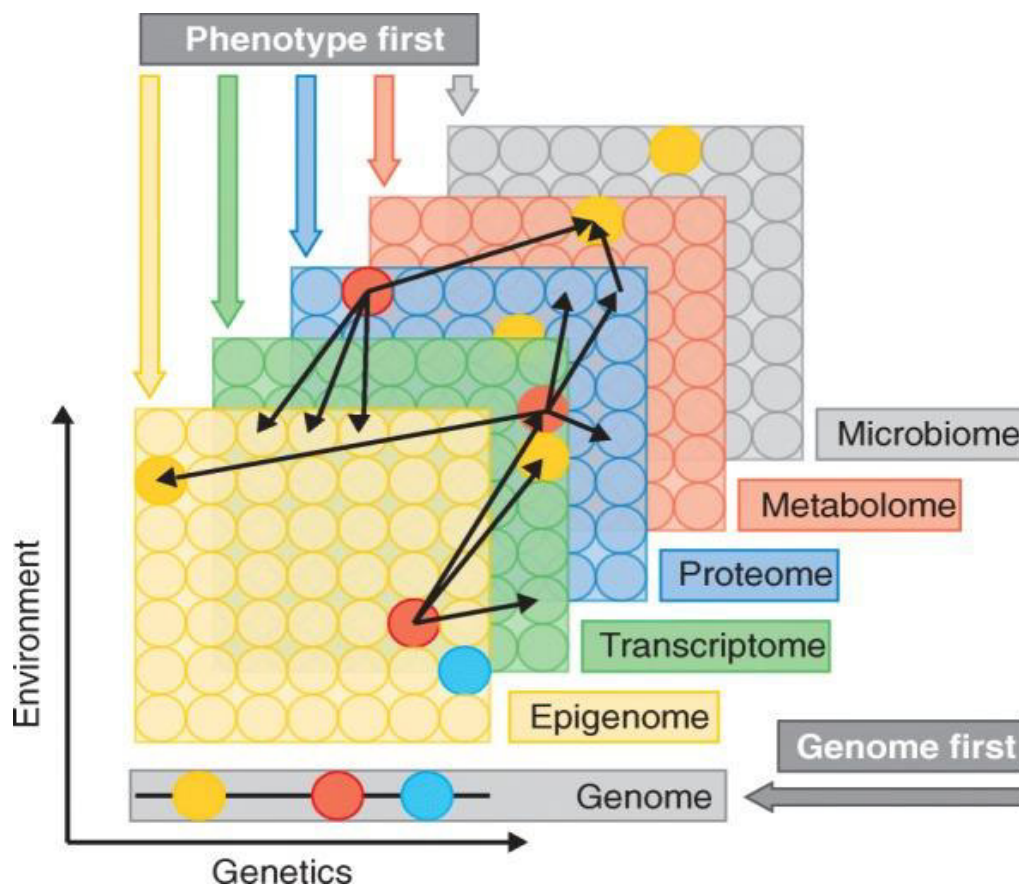


Figura 1. Relación de análisis de las ómicas. Tomada de Hasin *et al.*, 2017 (50).

2.3.1 Importancia de la aplicación de ómicas en el análisis de alimentos

El continuo cambio en el análisis y estudio de la industria alimentaria ha generado la creación y ampliación del término Alimentómica (“*Foodomics*” en inglés). Actualmente el término engloba una colección de enfoques integrando información desde algunas áreas donde el alimento (más específicamente la nutrición), técnicas analíticas avanzadas (ómicas, posiblemente en conjunto los estudios *in vitro*, *in vivo* y clínicos) y el análisis bioinformático (51).

El término “*foodomics*” abarca la genómica, transcriptómica, proteómica, peptidómica, metabolómica, lipidómica y tecnologías similares de alto rendimiento y resolución que ofrecen diversas oportunidades para evaluar la calidad e inocuidad de los alimentos y su origen, tanto animal o vegetal.

Comenzando desde la evaluación de la salud tanto de la planta como del animal como de los productores (52), pasando por la producción y monitoreo tanto de la

calidad e inocuidad del alimento, el proceso finaliza con el control de la calidad y autenticidad, y protección contra las adulteraciones en los correspondientes laboratorios de análisis (52–55).



Figura 2. Análisis de la Alimentómica desde la producción hasta el consumo. Modificado de Andjelkovic *et al.*, 2017 (56).

2.4 Transcriptómica

Es el estudio del transcriptoma, toda la información de los transcritos de ARN que son producidos por el genoma, bajo circunstancias completamente direccionadas o en una célula específica, usando métodos de alto rendimiento, tales como análisis de microarrays o metodología NGS (57).

Los objetivos clave de la transcriptómica son: catalogar todas las especies de transcripción, lo que incluye a los ARNm, los ARN no codificantes y los ARN cortos; para determinar la estructura transcripcional de los genes, en términos de sus sitios de inicio, extremos 5' y 3', patrones de empalme y otras modificaciones pos-transcripcionales; y para cuantificar los niveles de expresión cambiantes de cada transcripción durante el desarrollo de cualquier ser vivo y en diferentes

condiciones (58). La comparación de diferentes transcriptomas de una misma especie permite la identificación de genes que han sido diferencialmente expresados en distintas poblaciones celulares, o en respuesta a diferentes tratamientos (57).

La identificación de transcritos y la cuantificación de la expresión génica han pasado a tener diferentes actividades en la biología molecular a partir de la impartición de rol protagónico del ARN como el nexo entre el genoma y proteoma (58).

2.4.1 Tecnología para el análisis del transcriptoma

Se han desarrollado a lo largo del tiempo en estos últimos años diversas tecnologías para deducir y cuantificar el transcriptoma, incluyendo los enfoques en hibridación o en secuencias. Las tecnologías basadas en la hibridación comúnmente implican incubar ADN complementario marcado de forma fluorescente con microarrays hechos a medida u oligo microarrays comerciales de alta densidad. También se han diseñado microarrays especializados; por ejemplo, se pueden usar matrices con sondas que abarcan uniones exónicas para detectar y cuantificar distintas isoformas unidas (59).

Las tecnologías basadas en la hibridación son relativamente económicas y tienen un alto rendimiento. Sin embargo, así como tienen factores positivos, también tienen limitaciones, que incluyen: confiar en el conocimiento existente sobre la secuencia del genoma, altos niveles de fondo debido a la hibridación cruzada (60,61) ; y un rango dinámico limitado de detección debido al fondo y la saturación de señales. Además, la comparación de niveles de expresión en diferentes modelos experimentales es comúnmente compleja y puede requerir métodos de normalización difíciles.

A diferencia de los métodos de microarrays, los enfoques basados en secuencias determinan directamente la secuencia de ADNc. Inicialmente, se utilizó la secuenciación de Sanger de las librerías de ADNc o EST(62,63), pero esta tecnología posee un rendimiento bastante bajo, es costosa y generalmente no se realiza un análisis cuantitativo, es puramente cualitativo. Tecnologías digitales basadas en el método de Sanger fueron desarrolladas para superar dichas

limitaciones, las que incluían el análisis en serie de la expresión génica (SAGE, siglas debido a su descripción en inglés) y análisis del límite de la expresión génica (CAGE, siglas debido a su descripción en inglés) (64,65). A pesar de ser tecnologías de alto rendimiento, el hecho de que estén basadas en el método de Sanger, las sigue haciendo costosas y, además, porciones significativas de secuencias cortas no pueden ser mapeadas únicamente al genoma de referencia, siendo limitantes bastante significativas (58).

Recientemente, el desarrollo de nuevos métodos para el secuenciamiento de alto rendimiento del ARN permite tanto el mapeo como la cuantificación del transcriptoma. Este método denominado secuenciamiento del ARN (RNA sequencing = RNASeq), tiene claras ventajas sobre las tecnologías previamente descritas y ya viene aplicándose ampliamente en el mundo de la ciencia para el análisis de transcriptomas eucarióticos. Algunos ejemplos de especies a las que se le ha realizado tal análisis son: *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Arabidopsis thaliana* (66,67), ratón y células humanas.

La potencia del secuenciamiento del RNA se basa en el hecho de que dos aspectos fundamentales tales como el descubrimiento de secuencias y su cuantificación, pueden llevarse a cabo con la tecnología RNA-Seq (68).

2.4.1.1 Secuenciamiento del ARN (RNA sequencing)

Se define como RNA-Seq a la combinación de una metodología de secuenciamiento de alto rendimiento con enfoques computacionales y softwares para capturar y cuantificar transcritos presentes en una extracción de ARN a partir de un tejido o una célula (69). RNA-Seq tiene la aplicación de identificar genes dentro del genoma o cuáles genes se están expresando en ese determinado momento en el tiempo, y los recuentos de lectura pueden ser usados para modelar con precisión el nivel relativo de expresión génica. Esta metodología ha mejorado con creces en el tiempo con la finalidad de aumentar la precisión, el rendimiento y la longitud de lectura (70).

Los transcritos de RNA-Seq se transcriben de manera inversa en ADNc, y los adaptadores se ligan a cada extremo del ADNc. El secuenciamiento se puede

realizar de forma unidireccional (secuenciamiento de extremo único o según su denominación en inglés *single end sequencing*) o bidireccional (secuenciamiento de extremo pareado o *paired end sequencing*) y luego alinearla con una base de datos de genoma de referencia o ensamblarla para obtener transcripciones de *novo*, lo que evidenciaría un perfil de expresión de genoma completo (58), tal como se aprecia en la figura 3.

A diferencia de la tecnología de microarrays, que basa su análisis en los genes ya conocidos, RNA-Seq no depende de los datos del genoma existentes y puede seleccionar transcripciones nuevas y analizar su estructura, incluidas la resolución de un solo par de bases y los límites exónicos, lo cual es muy valioso al investigar los SNPs. haciéndolo útil para el genotipado y el análisis de ligamiento (58). Otra marcada diferencia con el análisis por microarrays es que para RNA-Seq solo hace falta una muestra de entrada de cantidades en nanogramos, mientras que para los microarrays se necesitan cantidades en microgramos, lo que permite una evaluación mucho más fina de las estructuras celulares, hasta el nivel de analizar una sola célula cuando se mezcla con la amplificación lineal del ADNc (71).

La adopción generalizada de RNA-seq se ha extendido más allá de la comunidad avocada netamente al estudio de la genómica y se ha convertido en una herramienta estándar usada por la comunidad de científicos en general dedicados al estudio de las ciencias de la vida (68).

RNA-Seq utiliza tecnologías de secuenciación profunda desarrolladas recientemente. En general, un conjunto de ARN (total o fraccionado, como poli (A)+) se convierte en una biblioteca de fragmentos de ADNc con adaptadores unidos a uno o ambos extremos. Cada molécula, con o sin amplificación se secuencia en una plataforma de alto rendimiento para obtener secuencias más cortas de un extremo (*single end*) o de ambos extremos (*paired end*). Las lecturas normalmente oscilan entre 30 a 400 pb, dependiendo esto de la tecnología de secuenciación empleada. En principio, se puede utilizar cualquier tecnología de secuenciamiento (72), siempre que sea de alto rendimiento para RNA-seq, entre estas destacan Illumina IG (66,67,73,74), SOLiD Applied Biosystems (75) y Roche 454 Life Science (76,77).

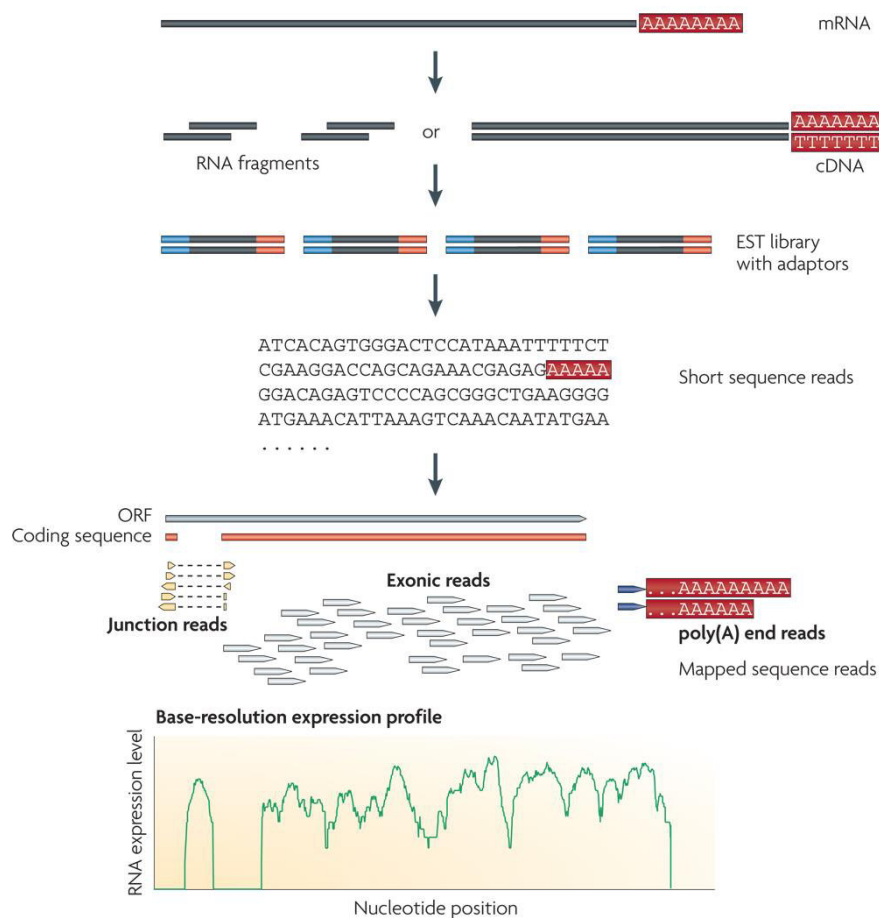


Figura 3. Experimento típico de RNA-Seq. Tomado de Zhong *et al.*, 2009 (58).

2.4.1.2 Tecnología aplicada para RNA-Seq: Secuenciación de Próxima Generación (NGS)

La era de la genómica comenzó en 1977 con el secuenciamiento del bacteriófago phiX174 por Fredrick Sanger (78). Las posteriores mejoras al método, hicieron de Sanger el enfoque de secuenciamiento dominante en el mundo entero, el cual fue empleado para el colosalmente ambicioso HGP (1990 - 2003), (79).

La necesidad de una tecnología de secuenciamiento de alto rendimiento se acrecentó debido a la iniciación de Proyecto del Genoma Humano en 1990 y su finalidad de secuenciar e interpretar 3.2 billones pares de base de nucleótidos que comprenden el genoma humano para posibles beneficios médicos (80).

Un mayor cambio ocurrió a inicios de este milenio, cuando los nuevos métodos de secuenciación de próxima generación (NGS, por sus siglas en inglés) que utilizan procesamiento masivamente paralelo, redujeron significativamente los costos a comparación del método de Sanger, así como el tiempo de secuenciación. Durante la década del 2010, WGS (según siglas en inglés de secuenciamiento de todo el genoma) comenzó a migrar de simples laboratorios de ensayo a prácticas de salud pública (81).

El método de Sanger (tecnología de primera generación) fue el método de secuenciación primordial entre 1975 y 2005. Este produce secuencias relativas de ADN de alta calidad (500 – 1000 bp), y ha sido largamente considerada la referencia estándar para el secuenciamiento de ADN. La aparición de la tecnología por piro secuenciamiento por 454 Life Sciences (Branford, CT, USA) en 2005 comenzó la era de la revolución NGS. Esta tecnología de alto rendimiento permitió la generación y detección de miles de millones de lecturas de secuencias cortas en una simple máquina de corrida sin la necesidad de realizar clonación. Desde entonces, han surgido muchas otras tecnologías NGS que generan lecturas cortas (50-400 pb) y largas (1-100 kb) (81).

Los diversos proveedores de plataformas tecnológicas de secuenciamiento han creado diferentes estrategias para diseñar las librerías de las secuencias en plantillas adecuadas, así como para identificar la señal la señal y finalmente leer la secuencia de ADN o ARN.

La bibliografía divide las tecnologías NGS en dos tipos (82,83). Se reconoce a la segunda generación de tecnologías, las cuales se refieren como las más recientes tecnologías de secuenciamiento desarrolladas en la atmósfera NGS después de la primera generación (84,85), estas metodologías se caracterizan por la necesidad de preparar bancos de secuenciamiento amplificados antes de empezar con el secuenciamiento de los clones de DNA amplificados (86) y luego está la tercera generación de tecnologías de secuenciamiento, recientemente introducidas al campo de la investigación de los datos en masa (87), en contraste con la segunda generación, estas tecnologías están clasificadas como secuenciamiento de una sola molécula debido a que pueden secuenciar una sola

molécula sin la necesidad de crear librerías de amplificación y son ampliamente capaces de lecturas largas en menor tiempo y a más bajo costo (88).

Actualmente, una considerable variedad de plataformas NGS están disponibles para el uso de la ciencia. Las más importantes cualidades de estas plataformas NGS, tanto la salida como longitud del fragmento, están descritas en la tabla N°5.

Las diversas plataformas NGS hacen uso de diferentes tecnologías de secuenciamiento. Los secuenciadores de Illumina realizan el secuenciamiento mediante síntesis de fluorescentes y terminadores reversibles. Los secuenciadores de ThermoFisher (Ion Torrent) usan secuenciamiento semiconductor, que mide las modificaciones de pH durante la incorporación de nucleótidos. Pacific Biosciences utiliza nucleótidos fluorescentes en su tecnología de tiempo real de una sola molécula (SMRT), y las plataformas Oxford Nanopore, que están incluidas en las tecnologías de tercera generación, usan la localización de corriente iónica, en la cual el ADN se desplaza a través de nano-poros, cambiando así la corriente de una manera que es específica para cada tipo de nucleótido (89,90).

Las tecnologías que detectan lecturas cortas actualmente en uso, son conocidas en conjunto como secuenciamiento paralelo masivo y son comúnmente denominadas tecnologías de secuenciamiento de segunda generación (91). Producen billones de secuencias de nucleótidos durante cada corrida, donde cada genoma es secuenciado repetidas veces en pequeñas piezas aleatorias para generar enormes bases de datos. Aunque cada plataforma posee una bioquímica y ensayos exclusivos, los flujos de trabajo describen similares pasos: (i) extracción de ADN o ARN dependiendo el tipo de análisis; preparación de la librería, que generalmente incluye el corte del ADN o ARN ya sea de forma mecánica o enzimática; (iii) preparación de la plantilla, ya sea por amplificación o PCR de emulsión, para finalmente realizar el secuenciamiento automatizado (92).

Las técnicas de NGS pueden ser clasificadas de forma general en aplicaciones para la investigación del genoma, transcriptoma y epigenoma. Los ensayos genómicos incluyen WGS, WES y un nuevo secuenciamiento de regiones

específicas objetivas con la finalidad de descubrir variantes asociadas con la función celular o el desarrollo de alguna enfermedad (93).

El análisis del transcriptoma basado en la tecnología NGS (RNA-seq) incluye el perfil de la expresión génica cuantitativa, el descubrimiento de nuevas secuencias transcritas o secuencias *de novo* cuando se realiza el comparativo con el genoma de referencia (64) y finalmente, la identificación de especies de ARN no codificante tales como miARN y lncARN (94,95).

Los métodos de análisis del epigenoma, están dirigidos a la estructura de la cromatina e incluyen DNase-seq (96), ATAC-seq (Buenrostro *et al.*), metilación del ADN (73), y la modificación de histonas a través de ChIP-seq (97). Actualmente se sincroniza el estudio de ambos debido a la estrecha relación que guardan en el silenciamiento y sobreexpresión de genes objetivos.

Tabla N°5. Principales Plataformas de Secuenciamiento usadas en la actualidad.

Grupo	Plataforma	Instrumento	Tipo de corrida	Tiempo de corrida	Longitud de lecturas (pb)	# de lecturas por corrida	Año
Primera Generación	ABI Sanger	3730xl	SE	7h	400-900	96	2002
Segunda Generación	Illumina	MiSeq	SE & PE	5-55h	300	25M (máximo)	2011
	Illumina	NextSeq	SE & PE	12-30h	150	400M (máximo)	2014
	Illumina	HiSeq	SE & PE	1-3.5 días	150	5B (máximo)	2012
	Ion Torrent (ThermoFisher)	Ion Proton	SE	<24h	200	60M-80M	2012
	Solid	5500 W	SE	6-7 días	75	3B	2011
Tercera Generación	PacBio	RS C2 XL	SE	30min-6h	4300	432	2011
	Oxford Nanopore	Min ION Mk	1D, 2D	1min-48h	9455	100	2015
*SE: single end; PE: paired end, pb: pares de base							

Fuente: Modificado de Deurenberg *et al.*, 2017 (98), Kchouk *et al.*, 2017 (88).

2.4.1.3 Tecnología de secuenciamiento Illumina/Solexa

La tecnología Illumina realiza el secuenciamiento por síntesis y es actualmente la tecnología más utilizada en el mundo NGS (88).

El proceso de secuenciamiento para ARN consta de dos etapas: preparación de la librería y el posterior secuenciamiento en la plataforma elegida (MiniSeq, NextSeq o HiSeq). Para la preparación de la librería, en primer punto, se aísla el mRNA utilizando perlas de oligo dT. El mRNA es usado para hacer la primera y segunda hebra de cADN y estas moléculas de doble hebra de cADN son enzimáticamente fragmentadas. Las terminaciones de estas moléculas son reparadas y un nucleótido A es añadido. Las muestras ligadas son enriquecidas mediante amplificación mediante cebadores específicos adaptadores y purificados para el secuenciamiento posterior (99).

2.4.2 Análisis Bioinformático

Muchas variantes de protocolos y análisis para RNA-Seq se han propuesto, representando un desafío para los nuevos usuarios en identificar cuál sería el adecuado para su modelo de estudio (68).

Actualmente, no existe solo una secuencia de software, diseñada para la diversidad de aplicaciones y diferentes escenarios en los que RNA-seq es aplicado. Generalmente, los científicos suelen diseñar un pipeline exclusivo y adoptar diferentes estrategias de análisis dependiendo del organismo que es materia de estudio y los objetivos específicos del mismo (68).

2.4.2.1 Control de calidad

El control de calidad de las lecturas sin procesar (raw reads), incluye el análisis de calidad de la secuencia, contenido de GC, la presencia de adaptadores, sobre-representados en k-mers y lecturas duplicadas para detectar errores de secuenciamiento, artificios de PCR o contaminaciones. Se recomienda que los valores atípicos con más del 30% de disconformidad sean descartados. FastQC (100) es una herramienta bastante utilizada y solicitada para llevar a cabo el análisis de control de calidad para lecturas (reads) provenientes de la plataforma Illumina. Entre las principales funciones de esta herramienta están: importar data

desde documentos BAM, SAM o FastQ (cualquier variante), otorgar una rápida visión panorámica para decirle al investigador en qué áreas se podrían presentar problemas, resumir en gráficos y tablas para evaluar la gran cantidad de data en poco tiempo, exportar los resultados en un reporte permanente HTML (101). Mientras NGSQC puede aplicarse para el control de lecturas emitidas por cualquier plataforma (102).

2.4.2.2 Alineamiento de lecturas

Las lecturas son regularmente mapeadas para, ya sea, un genoma o transcriptoma. Un transcendental parámetro para la calidad del mapeo es el porcentaje de lecturas, el cual es un indicador de la precisión de las secuencias y la presencia de ADN contaminante. Por ejemplo, se espera entre 70 a 90% de lecturas regulares de RNA-seq para mapear sobre el genoma del ser humano (dependiendo sustancialmente del software de mapeo utilizado) (103).

Para poder mapear las lecturas hay dos posibilidades: hacerlo con un genoma o transcriptoma de referencia o sin estos. Cuando un genoma de referencia está disponible, el análisis RNA-seq involucra el mapeo de las lecturas sobre el transcriptoma o genoma de referencia para poder evidenciar qué transcritos se están expresando. Cuando no se posee un genoma de referencia del organismo en estudio, en primer lugar, se deben ensamblar las lecturas en grandes *contigs* y luego tratar estas como el transcriptoma expresado para el cual las lecturas son mapeadas nuevamente para la cuantificación (68).

Uno de los software de mapeo para RNA-seq más famosos es TopHat, el cual sigue una estrategia de análisis de dos pasos en los cuales las lecturas sin cortar, en primer lugar, son mapeadas para localizar exones, en segundo lugar, las lecturas que no han sido mapeadas son divididas y alineadas independientemente para identificar uniones de exones (104,105), lo que es graficado de clara forma en la figura 4.

Algunos otros softwares de mapeo, al ser mejorados, se han encargado de detectar objetivos específicos: GSNA (106), PALMapper (107), MapSplice (108) para SNPs; STAR (103) y MapSplice (108) para uniones de empalme no

canónicas; GEM (109) para lograr mapeos ultra rápidos o STAR (103), también, para mapear lecturas largas.

Hay parámetros sumamente fundamentales a considerar durante el mapeo, como el trenzado de la librería de RNA-seq, el número de mismatches que se pueden aceptar, la longitud y el tipo de lecturas (SE o PE), y la longitud de los fragmentos secuenciados (68).

2.4.2.3 Cuantificación de Transcritos

Una vez los valores de la cuantificación han sido calculados, estos deben ser revisados para evaluar el contenido GC y el sesgo de la longitud del gen, entonces esa corrección de los métodos de normalización puede ser aplicado si es necesario (68).

La aplicación más común de un experimento RNA-Seq es deducir, tanto la expresión génica como de los transcritos. El número de lecturas que se mapean para cada lectura de transcrito es la aplicación habitual de RNA-Seq (110).

Agregar lecturas sin procesar utilizando software tales como HTSeq-count (111) o featureCount (112) es el más simple enfoque que se le da a la etapa de cuantificación de transcritos. Para el análisis de cuantificación de los niveles génicos se utiliza un formato de transferencia de genes o GTF (Gene Transfer Format) (113), que usualmente tiene por contenido las coordenadas de genes y exones, y frecuentemente descarta las lecturas múltiples.

Dentro de este contexto de cuantificación, los valores de expresión de RNA-Seq son reportados en las siguientes unidades: RPKM (lecturas por kilobase de exón modelo por un millón de lecturas), FPKM (fragmentos por kilobase de exón modelo por un millón de lecturas mapeadas) y TPM (transcritos por millón). El primero sigue un método dentro de la muestra de normalización que se encarga de remover los efectos de la longitud característica y el tamaño de la librería (68).

Programas, tales como Cufflinks que infieren la longitud del gen a partir de la data pueden hallar diferencias bastante significativas en la longitud del gen evaluados entre muestras que no pueden ser ignoradas. También, Cufflinks, estima expresión de los transcritos usando un enfoque de maximización – expectación,

obtenidos a partir de software como TopHat, como se visualiza en la figura 4. Esta herramienta ha sido diseñada para aprovechar lecturas PE (pair end) y puede usar información a partir de formatos GTF para hallar los transcritos expresados, o poder inferir transcritos *de novo* a partir de una data única mapeada (114).

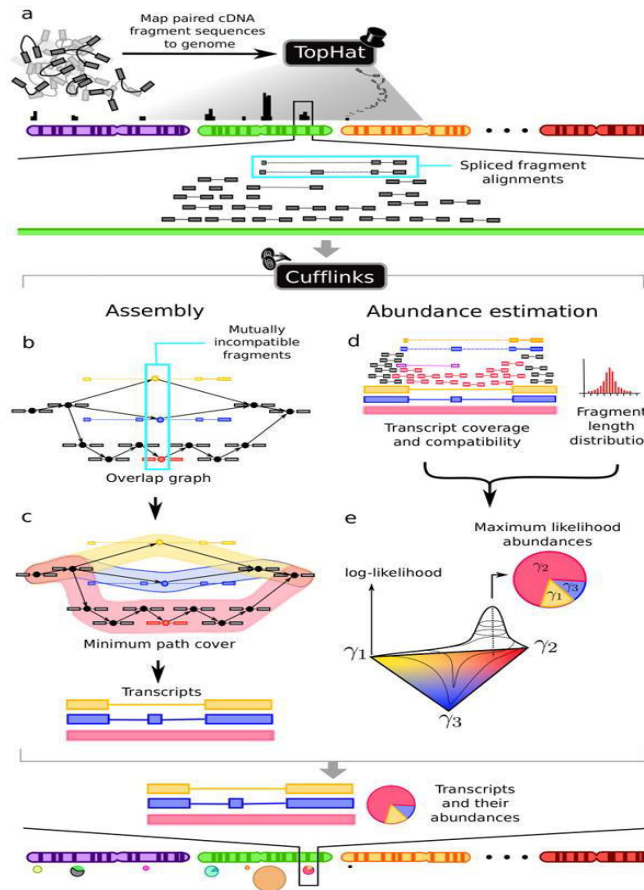


Figura 4. Mecanismos de acción de lo aplicado por los software TopHat y Cufflinks. Tomada de Trapnell *et al.*, 2010 (114).

*Descripción general de Cufflinks. El algoritmo toma como entrada las secuencias de fragmentos de cDNA que (a) están alineadas con el genoma por un software capaz de producir alineamientos empalmados, como TopHat. Con PE RNA-Seq, Cufflinks trata cada par de fragmentos como una sola alineación. (b) Ensamblaje de los transcritos en donde se descartan múltiples fragmentos que no son compatibles. (c) Se obtienen fragmentos que cubren una mínima parte de las rutas para empalmarlos con el genoma de referencia. (d) La capacidad de Cufflinks para estimar la cantidad de transcritos similares que cubren una misma región del genoma de referencia basados en su compatibilidad. Finalmente se consigue estadísticamente a probabilidad máxima de la abundancia de cada transcrito.

Habitualmente encontramos otros algoritmos que cuantifican la expresión a partir del mapeado del transcriptoma. Estos algoritmos incluyen a, RSEM (RNA-Seq por Maximización – Expectación) (115), eXpress (116), Sailfish (110) y kallisto (117) entre otros. Todos estos softwares permiten localizar lecturas multimapeadas entre valores transcritos normalizados dentro de una muestra.

2.4.2.4 Expresión Diferencial Génica

Este análisis necesita los valores de expresión génica entre muestras deban ser comparadas. RPKM, FPKM y TPM normalizan el factor más importante para comparar muestras, que es la profundidad de la secuencia, ya sea solo tomando en cuenta el número de transcripciones o directamente, que pueden diferir significativamente entre las muestras (68).

Estos enfoques se sustentan en métodos normalizados estadísticamente que están basados en totales o conteos efectivos, y tienden a realizar vagamente cuando las muestras poseen distribuciones heterogéneas de transcritos. Los softwares que basan su análisis de expresión diferencial son TMM (118), DESeq (119), PoissonSeq (120) y UpperQuartile (121). El paquete estadístico NOISeq R, por su parte, contiene una amplia variedad de gráficos de diagnóstico para identificar fuentes de sesgo en la data de RNA-Seq y efectuar adecuadamente procedimientos de normalización en cada caso. Finalmente, a pesar de esta normalización específica de la muestra, la presencia de las grandes cantidades de información, aún pueden estar presentes en los datos (68).

Otros métodos, como el popular edge R (122), toma como partida recuentos de lectura e introduce posibles sesgos dentro del modelo estadístico para llevar a cabo una normalización integrada, así como un análisis de expresión diferencial. En otros programas, la expresión diferencial necesita de data que haya sido previamente normalizada para remover todo lo relacionado a sesgos. DESeq2, así como edge R, utiliza el binomio negativo como referencia de distribución y provee su propio enfoque de normalización (119,123).

Entre todas las herramientas para este análisis, lima ha demostrado realizar de la mejor forma bajo diferentes circunstancias y ha sido el más rápido en analizar los datos (124–126). Edge R y DESeq materializan parecido dentro del ranking de

genes, pero frecuentemente son relativamente conservativos o muy liberales, respectivamente, en control de FDR (125,127,128).

Mientras algunos softwares estadísticos para el análisis de expresión diferencial pueden realizar la comparación por parejas, otras como edgeR (122), limma-voom (129), DESeq (119), DESeq2 (123) y maSigPro (120) pueden hacer múltiples comparaciones, incluyendo diferentes covariantes.

Otro software que suele ser utilizado para este análisis diferencial, es Cuffdiff, del paquete de Cufflinks. Este software tiene como labor principal identificar a un nivel transcripcional o post-transcripcional, genes que tienen diferencia en su regulación. Estos resultados son reportados como un grupo de documentos que pueden ser mostrados en el programa de preferencia del usuario (130).

III. MATERIALES Y MÉTODOS

3.1 Lugar de realización

La elaboración de esta tesis se ha llevado a cabo en el laboratorio del Grupo de Investigación GENOBIC, perteneciente al Departamento de Bioquímica de la Facultad de Farmacia y Bioquímica de la Universidad Nacional Mayor de San Marcos.

3.2. Muestra

Se utilizaron 02 muestras de tejido adiposo de la parte inferior del tórax (pechuga) correspondientes a 02 pollos de engorde beneficiados.

Los criterios para tomar las muestras en dos lugares fueron los siguientes: frescura del producto y condiciones de higiene-saneamiento con marcada diferencia relacionada a su procedencia, con la finalidad de realizar el análisis transcriptómico comparativo.

El primer lugar, el Centro de Acopio de San Luis (CA001), con condiciones de inocuidad deficientes: infraestructura inadecuada (piso de tierra), contenedores de plástico sin higiene, contenedor de desechos sin tapa, tablas de madera para manipulación del producto totalmente desgastadas y sucias. Es en este mismo lugar donde se realiza el faenamiento del animal. Todo esto va en contra de lo establecido por SENASA en el Reglamento del Sistema Sanitario Avícola (131). Se infiere que a ese lugar llegan pollos con deficiencias nutricionales y criados en condiciones inadecuadas.

El segundo lugar de donde se obtuvo la muestra fue el Mercado Modelo de Bellavista (MB001). Las condiciones son totalmente distintas. Los manipuladores cuentan con la indumentaria adecuada, las instalaciones son de mayólicas de color claro, lo que denota la limpieza del lugar inmediatamente, el producto también se comercializa fresco y se conserva a temperatura de refrigeración (de 0°C a 4°C) hasta el momento de su venta, aunque también se tienen algunos de los productos más frescos expuestos a la intemperie, lo que termina siendo una mala praxis en este lugar. Se infiere que a ese lugar llegan los pollos con una adecuada nutrición y criados en condiciones apropiadas.

En ese contexto, se consideran las posibles condiciones previas a partir de los puestos de venta seleccionados. En ambos casos, estas podrían tener impacto en la expresión génica por lo que se propone ese análisis en esta investigación.

3.2.1 Toma de Muestra

Las muestras fueron tomadas a partir de los criterios descritos en el punto anterior. Se tomaron en una bolsa de polietileno de primer uso en ambos puntos y se conservaron en un contenedor (de tecnopor) a 4°C para poder transportarlas hasta el laboratorio de GENOBIDC, donde fueron procesadas.

3.2 Análisis Molecular

El tratamiento del tejido adiposo de *Gallus gallus* se realizó para la extracción del ARN total para el secuenciamiento masivo.

3.2.1 Tratamiento de la Muestra

En ambos casos, al tomar la muestra, se seleccionó la parte del pecho (parte conocida como pechuga) completo y en el laboratorio se realizó la extracción del tejido adiposo alojado en dicha porción de carne con un bisturí de primer uso. Se tomó una porción de tejido adiposo de cada pechuga según procedencia (entre 300 – 400 mg) para la extracción del ARN total con el reactivo trizol (procedimiento completo en los anexos). Para confirmar la extracción se realizó la electroforesis en gel de agarosa al 1.5%.

Aparte, para tener un apoyo o muestra de reserva en caso el ARN total no cumpliera con las condiciones requeridas para el secuenciamiento, se enviaron dos tubos con tejido adiposo (entre 400 – 600 mg) en tubos eppendorf de 1.5 suspendidos en 1 ml de reactivo RNA later, para que se pueda conservar el ARN del tejido, a temperatura de refrigeración (4°C).

3.2.2 Extracción de ARN total

Las muestras de ARN, obtenidas en el laboratorio GENOBIDC, enviadas para ser procesadas en la empresa MacroGen – Korea, fueron las que finalmente se utilizaron para proceder con el secuenciamiento RNA-seq. Previamente pasaron por un riguroso control de calidad. En primer lugar, se verificó la extracción efectiva del ARN. Se corrió lo extraído en gel de agarosa teñido con el reactivo fluorescente no mutagénico (CSL-RUNSAFE), leído al transiluminador LED de luz azul como se puede apreciar en la figura 8. Posteriormente, se determinó la cantidad de ARN en la plataforma Picogreen usando el método fluorométrico, mientras que el número de integridad del RNA (RIN) fue obtenido mediante el Bioanalizador Agilent Technologies 2100. Para efectos de obtener un ARN sin contaminantes, el ADN es digerido por las DNasas que contiene el reactivo utilizado en el protocolo y además, previa elaboración de la librería, algún remanente de este es removido durante la purificación del ARN mensajero.

Se describe el protocolo de extracción en Anexos.3.

3.2.3 Secuenciamiento del Transcriptoma (Protocolo Illumina)

Los ARNs extraídos cumplieron con los requerimientos necesarios (Número de integridad de ARN y concentración de ARN) fueron sometidas a la preparación de la librería (RNA-seq). Para poder realizar este procedimiento, el ARN total tuvo que ser convertido a ADN complementario, pasando por el protocolo indicado por la guía de preparación de la muestra: “TruSeq Stranded Total RNA” para la preparación de la librería.

La plataforma Next Generation Sequencing utilizada para el secuenciamiento fue el NovaSeq 6000. El sistema NovaSeq 6000 representa la plataforma de secuenciación de la tecnología de Illumina de alto rendimiento, el instrumento, se basa en la secuenciación por síntesis (*sequencing by synthesis*). El método Truseq stranded de ARN total con el kit Ribo-Zero fue empleado para la preparación de la librería tipo *Paired-end* (PE).

En general, el procedimiento para el secuenciamiento por la tecnología NGS de Illumina incluye 4 pasos:

- 1) **Preparación de la muestra:** para la construcción de la librería, el ARN es extraído de la muestra. Después de aplicar el control de calidad (quality control, QC), las muestras que califican, proceden para la construcción de la librería.
- 2) **Construcción de la librería:** la librería es preparada a partir de la fragmentación aleatoria de la muestra de ADN complementario, seguido por la ligación 5' y 3'. Alternativamente, la “tagmentación” combina las reacciones de fragmentación y ligación en un solo paso que incrementa enormemente la eficiencia del proceso de la preparación de la librería. Los fragmentos ligados al adaptador se amplifican por PCR y se purifican en gel.
- 3) **Secuenciamiento:** para la generación de agrupamientos, la biblioteca se carga en una celda de flujo donde los fragmentos se capturan en un campo de oligos unidos a la superficie y que son complementarios a los adaptadores de la librería. Luego, cada fragmento se amplifica por grupos. Cuando se completa la generación del clúster, las plantillas están listas para la secuenciación propiamente dicha en el equipo.
- 4) **Datos sin procesar (*raw data*):** la data resultado del secuenciamiento es procesada por el software y genera un archivo de tipo FastQ.

3.3 Análisis Bioinformático

La *raw data* es recibida en formato FastQ (figura 5). Por cada muestra se generan dos archivos ya que el tipo de corrida utilizado para la librería es *paired – end*, en la que se obtiene lecturas de las dos terminaciones, tanto en el sentido de ‘forward’ como en el sentido ‘reverse’. Este tipo de librería es más costoso, pero brinda mucha mayor información que el tipo de librería *single-end*; da con posiciones adicionales en relación al genoma y es usado para descifrar ensamblajes *de novo*.

En primer lugar, el primer tipo de secuencias que surge de la plataforma digital “next generation sequencing”, es el llamado diagrama de flujo o cromatograma. Estos tipos de secuencia están representados por una serie de picos a lo largo del tiempo donde cada pico es la intensidad de la señal y el tiempo es el orden de las

bases dentro de la secuencia de ADN. Para transformar estos picos se debe usar una “algoritmo de llamada de base” para designar un nucleótido a cada pico presente en el archivo en sin procesar (132).

El “algoritmo de llamada de base” más común es Phred”(133); la calidad de cada nucleótido dentro de una secuencia de ADN se expresa normalmente como “puntaje de calidad de Phred”. El puntaje de calidad, denominado Q , asignado a una base, es proporcional a su probabilidad de error, expresado como P . Un puntaje de calidad de Phred de 30, por ejemplo, corresponde a una probabilidad de error de 0.1% (134)

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGCTTTTTTGTGGAAACGAAAGGGTTTGAATTTCAAACCTTTTCGGTTTCCAACCTTCCA
AAGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""7F@71,'";C?,B;?6B;:EA1EA1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D
@/=<?7=9<2A8==

@ [title and optional description]
sequence line(s)
+ [optional repeat of title line]
quality line(s)
```

Figura 5. Ejemplo de lectura obtenida de formato FastQ. Tomada de Bacci, 2015 (132).

**La primera línea contiene el ID de la secuencia y es precedida por un “@”, seguido por un identificador de la secuencia. La segunda línea contiene la secuencia de ADN como una repetición de cuatro caracteres, uno por cada nucleótido (“A” para adenina, “C” para citosina, “T” para timina y “G” para guanina). La tercera línea comienza con el carácter “+” que puede ser secundado por una repetición del ID de la secuencia (la misma presentada en la primera línea), o no. Para finalizar, la cuarta línea posee los valores de calidad, y tiene que contener el mismo número de símbolos como letras en la secuencia. Esta línea se codifica por un simple carácter ASCII para cada base. Cada carácter ASCII puede ser simbolizado como un número entero de 0 a 128 (132). Sin embargo, la puntuación de calidad de un nucleótido puede variar de 0 a 40 Phred (41 para las plataformas más recientes de Illumina).*

En detalles, la *raw data* generada en formato FastQ, fue analizada en la plataforma web www.usegalaxy.org donde encontramos diferentes software libres para el análisis RNA-Seq y programas bioestadísticos como el paquete R estudio, considerando las siguientes etapas:

3.3.1 Control de Calidad de las lecturas secuenciadas

Los parámetros de calidad de los transcritos se evaluaron con el programa FASTQC (101), el cual se encarga de expresar una serie de gráficos (figura 6), entre las que sobresalen:

**Estadística Básica:* el programa produce una simple composición estadística de los archivos analizados. En este se puede encontrar información como la longitud de las secuencias, el total de las secuencias procesadas, codificación ASCII, entre otros (Figura 6 – A).

**Calidad de la Secuencia por Base:* para cada posición de lectura se dibuja un gráfico de tipo cajas. La línea central roja en cada caja es el valor de la mediana, la caja amarilla representa el rango de los inter-cuartiles (25-75%), los bigotes arriba y debajo de la caja amarilla simbolizan el 10% y 90% de puntos, y finalmente la línea azul simboliza la calidad (Figura 6 – B).

**Contenido por base de la secuencia:* este gráfico señala la proporción porcentual de cada base en cuanto a su posición en el archivo para el cual cada una de las cuatro bases de ADN ha sido expresada por colores. Permite identificar contaminación de la librería por presencia de marcados sesgos que cambian en las bases (Figura 6 – C).

**Duplicado de secuencias:* este módulo cuenta el grado de duplicación de cada secuencia en el conjunto y crea un gráfico que muestra el número relativo de secuencias con diferentes grados de duplicación. Un bajo nivel de duplicación puede ser señal de un alto nivel de cobertura en la secuencia objetivo (Figura 6 – D).

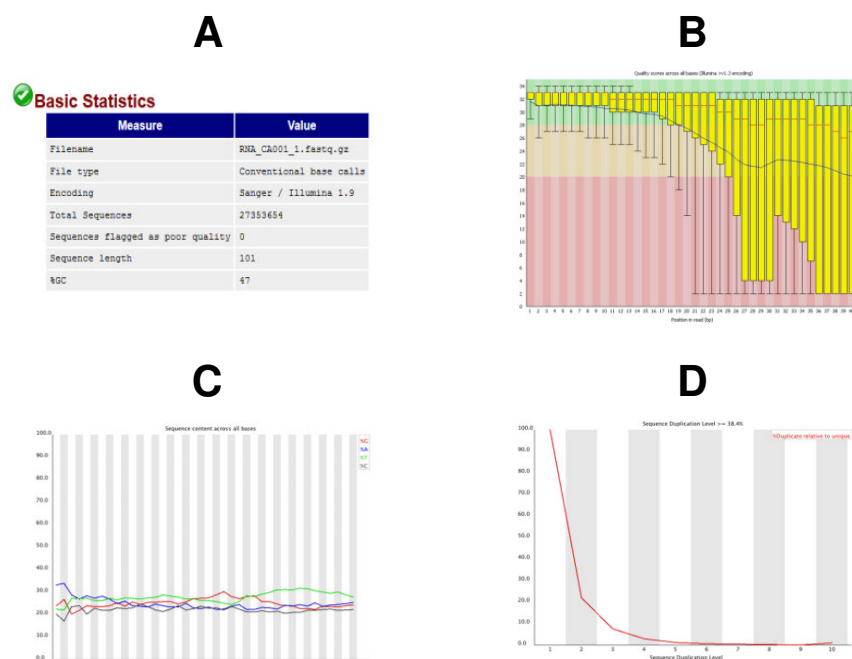


Figura 6. Gráficos del Control de calidad de la *Raw Data*, generados por el programa FastQC. Tomado de Babraham bioinformatics, 2010 (101).

3.3.2 Alineamiento de las lecturas: TopHat

Las lecturas secuenciadas son comúnmente mapeadas ya sea en relación a un genoma o un transcriptoma. Este paso del análisis se llevó a cabo con el software TopHat. Este software se encarga de encontrar uniones mediante el mapeo de lecturas de la secuencia en relación al de referencia en dos fases. En este estudio se utilizó como genoma de referencia el Gal.gal 4.0, disponible en la plataforma www.usegalaxy.org.

En la primera fase, el flujo asigna todas las lecturas en relación al genoma de referencia usando Bowtie. Todas las lecturas que no se asignan al genoma se reservan como "lecturas no asignadas inicialmente" o IUM reads (según siglas en inglés). El software selecciona primordialmente las lecturas con la terminación 5', debido a que poseen menor rango de errores dentro del secuenciamiento que aquellas con terminación 3' (135). TopHat permite que Bowtie reporte más de un alineamiento por lectura (por defecto son máximo 10), y elimina todos los alineamientos para lecturas que poseen más de estos 10, denominándose lecturas de baja complejidad, tampoco incluidas dentro de las IUM reads (48).

Luego TopHat ensambla las lecturas mapeadas usando el módulo ensamblador en Maq (136).

Se cargaron los archivos FastQ forward y reverse de las dos muestras (CA001 y MB001) en la plataforma virtual y se realizó el análisis en la página web Galaxy (<https://usegalaxy.org>) según las siguientes condiciones:

- a) **Distancia interna media entre pares: 300.** Script: `-r/--mate-inner-dist`; Esta es la distancia interna (media) esperada entre pares de parejas.
- b) **Desviación estándar para distancia entre pares de parejas: 20.** Script: `-mate-std-dev`; La desviación estándar para la distribución de distancias entre pares de parejas. El valor predeterminado es 20bp.
- c) **Genoma de referencia:** Chicken (*Gallus gallus*): galGal4
- d) **Máxima distancia de realineamiento: 1000.** Script: `--read-realign-edit-dist`; Esta opción indica a TopHat que vuelva a alinear las lecturas cuya distancia de edición de una alineación obtenida en un paso de mapeo anterior sea superior o igual a este valor de opción.
- e) **Distancia máxima de edición: 2.** Script: `--read-edit-dist`; Las alineaciones de lectura final que tengan más de estas distancias de edición se descartan.
- f) **Número final de fallas (mismatches): 2.** Script: `--read-mismatches`; Las alineaciones finales de lectura que tengan más que este número de desajustes se descartan.
- g) **Longitud de anclaje: 8.** Script: `-a / - min-anchor-length`; TopHat informará las uniones abarcadas por lecturas con al menos esta cantidad de bases en cada lado de la unión. Debe ser al menos 3 y el valor predeterminado es 8.
- h) **Número máximo de desajustes que pueden aparecer en la región de anclaje de la alineación empalmada: 0.** Script: `-m/--splice-mismatches`; El valor predeterminado es 0.
- i) **Número máximo de desajustes que pueden aparecer en la región de anclaje de la alineación empalmada: 70.** Script: `-i / - min-intron-length`; TopHat ignorará los pares de dadores / aceptadores más cercanos que estas muchas bases separadas. El valor predeterminado es 70.
- j) **Longitud máxima del intrón: 500000.** Script: `-l / - max-intron-length`; Al buscar uniones ab initio, TopHat ignorará los pares de donantes / aceptadores

más lejos que estas muchas bases de diferencia, excepto cuando dicho par esté respaldado por una alineación de segmento dividido de una lectura larga.

k) Longitud máxima de inserción: 3.

l) Longitud máxima de delección: 3.

m) Número máximo de alineaciones a ser permitidas: 20. Script: `-g / - max-multihits`; Indica a TopHat que permita hasta este número de alineaciones a la referencia para una lectura determinada, y elige las alineaciones según sus puntuaciones de alineación si hay más de este número.

n) Longitud mínima de intrón que se puede encontrar durante la búsqueda de segmento dividido (predeterminado): 50.

o) Longitud máxima de intrón que se puede encontrar durante la búsqueda de segmento dividido (predeterminado): 500000. Script: `--max-segment-intron`; La longitud máxima de intrón que se puede encontrar durante la búsqueda de segmento dividido.

p) Número de desajustes permitidos en cada alineación de segmento para lecturas asignadas de forma independiente: 2. Script: `--segment-mismatches`; Los segmentos de lectura se asignan de forma independiente, lo que permite una gran cantidad de desajustes en la alineación de cada segmento.

q) Longitud mínima de los segmentos de lectura: 25. Script: `--segment-length`; Cada lectura se corta en segmentos, cada uno por lo menos tan largo. Estos segmentos se asignan de forma independiente.

3.3.3 Ensamblaje de las lecturas: Cufflinks

Para poder efectuar el ensamblaje se utilizó el software libre Cufflinks dentro de la plataforma web Galaxy. Este se encarga de ensamblar los transcritos individuales a partir de las lecturas de RNA-seq que han sido alineadas en relación al genoma de referencia por el software TopHat. A causa de que una muestra puede poseer lecturas de múltiples variantes de empalme para un gen determinado, Cufflinks está preparado para deducir la estructura de empalme (*splicing*) de cada gen (130).

Luego de la fase de ensamblaje, el software cuantifica los niveles de expresión de cada transfrag en la muestra. Este término, transfrag, responde a una región genómica con actividad transcripcional de la cual se desconoce la estructura de

su transcrito precisamente. Diferentes transfrags pueden pertenecer al mismo transcrito, pero no hay información de conexión determinada. Debido a que estos transfrags se presentan en una cantidad muy baja en la librería en comparación con los transcritos empalmados (spliced transcripts), el software los excluye automáticamente (130).

Las condiciones bajo las cuales se llevó a cabo el análisis fueron las siguientes:

- a) **Longitud máxima de intrón: 300000.** Ignora los alineamientos con intrones más largos que estos.
- b) **Mínima cantidad de isoformas: 0.1.** Suprime transcritos por debajo de este nivel de abundancia.
- c) **Fracción de Pre mRNA: 0.15.** Suprime transcritos interintrónicos por debajo de este nivel.
- d) **Distancia media interior: 45pb.** Esta es la distancia interna (media) esperada entre pares de parejas.
- e) **Valor alfa para la prueba binomial utilizada durante la filtración de alineación de empalmes positivos falsos: 0.001**
- f) **Número máximo de fragmentos por locus: 1000000.** Establece el número máximo de fragmentos que un locus puede tener antes de ser omitido.
- g) **Tamaño mínimo de intrón permitido: 50pb.**
- h) **Cobertura promedio mínima requerida para intentar el recorte 3prima: 10**

Este software genera 4 archivos: 1) expresión génica, 2) transcritos ensamblados, 3) transcritos omitidos y 4) expresión de transcritos. Un archivo de cada uno por cada muestra (8 archivos en total).

3.3.4 Análisis de la expresión génica diferencial: Cuffdiff

Para poder realizar el análisis de la expresión diferencial se utilizó el software Cuffdiff, disponible en la plataforma virtual Galaxy. Este se encarga de calcular la expresión en dos o más muestras y probar la relevancia estadística de cambio detectado en la expresión entre dichas muestras (upregulated o sobre-regulados y downregulated o sub-regulados) (130).

Un documento con una tabulación simple que puede ser visualizada con cualquier programa que contenga hojas de cálculo (como Microsoft Excel), las diferencias y cambios en los niveles de transcripción son reportados por Cuffdiff (130).

El estadístico que utiliza este software para poder emitir los resultados en sus diferentes archivos es el método de dispersión ciega (Blind Dispersion Method); para así calcular el p valor y p valor ajustado. En este método todas las muestras son tratadas como réplicas de una sola condición. Otro punto destacable de este método para este trabajo, es que trabaja bastante bien cuando se espera de que las muestras no tengan demasiados genes diferencialmente expresados. También es recomendable para experimentos en los que se trabaja con una sola réplica biológica para las diferentes muestras, como en esta tesis (137).

Este software reporta el análisis diferencial de las muestras en cuestión contenido con diferentes documentos de salida como se muestra en la figura 7, en ambos casos, tanto desde la condición A como la condición B, provienen los archivos de los transcritos ensamblados (BAM), que se generaron a partir del software Cufflinks independientemente. Los parámetros que se utilizaron para obtener dicho archivo se describen en la tabla N°6.

Tabla N°6. Parámetros utilizados por el software Cuffdiff para el análisis diferencial génico.

PARÁMETRO	INFORMACIÓN
Transcritos	Transcritos ensamblados de Cufflinks CA001 y MB001
Tipo de archivo cargado	BAM
Muestra control	MB001
Muestra experimental	CA001
Método de Normalización de la Librería	Geométrico
Método de estimación de la dispersión	Agrupado
Tasa de descubrimientos falsos	0.05
Conteo mínimo de alineamiento	10

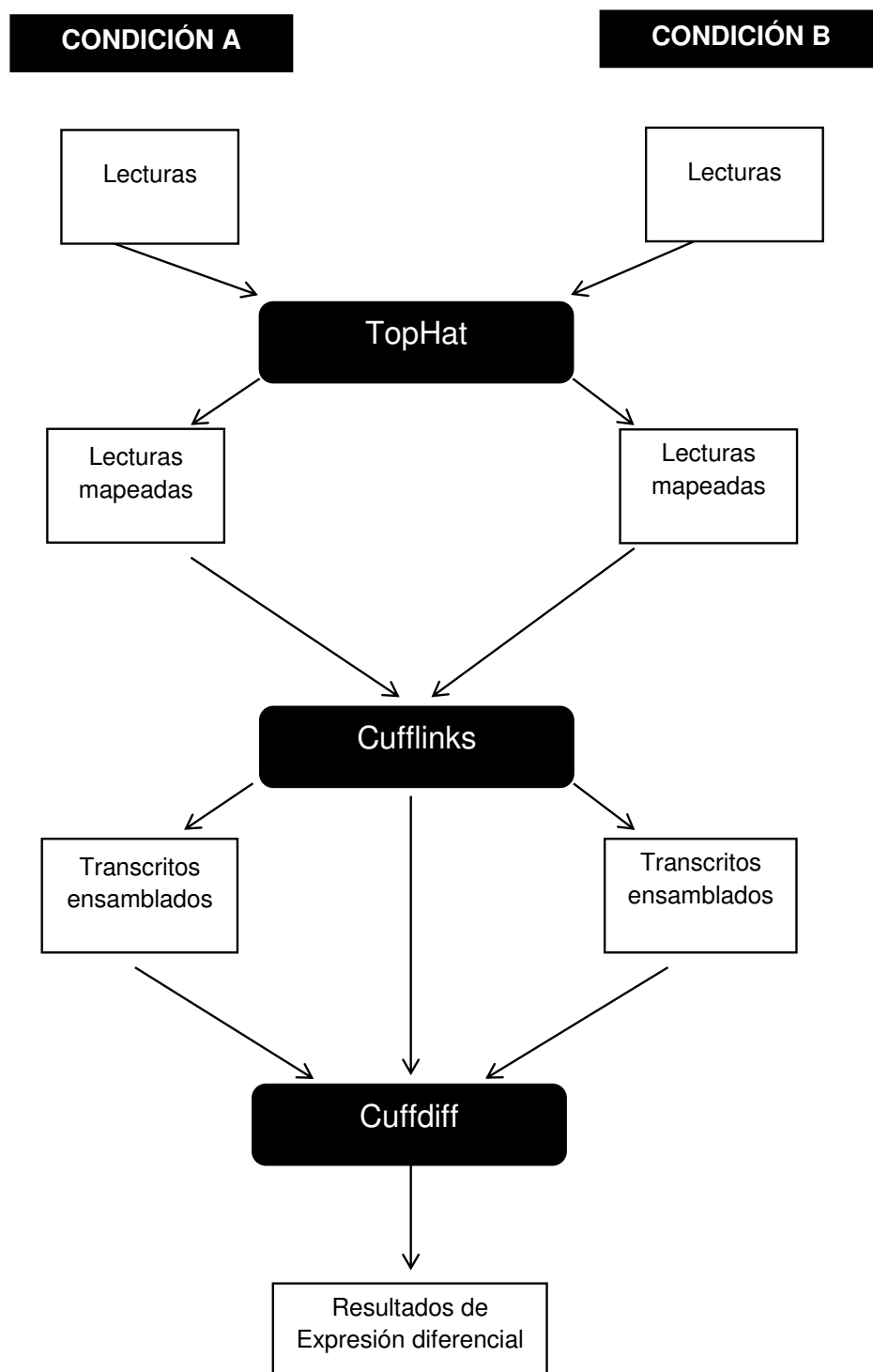


Figura 7. Esquema de la información que cada software del protocolo “Tuxedo” (tradicional) se encarga de procesar bajo dos condiciones que pueden ser control y experimental.

Modificado de Trapnell *et al.*, 2012 (130)

3.3.5 Análisis de rutas las metabólicas y funciones fisiológicas de los genes diferencialmente expresados (upregulated y downregulated): KEGG

Uno de los archivos que genera Cuffdiff contiene el nombre o denominación de los genes. En relación a ello, se realiza la búsqueda en la Enciclopedia de Kyoto de Genes y Genomas (KEGG según siglas en inglés) y la búsqueda de las rutas metabólicas en las que participan. Esta búsqueda se realiza a partir de los genes expresados diferencialmente (downregulated o sub-regulados y upregulated o sobre-regulados) de los genes.

IV. RESULTADOS

4.1 Toma de muestra

Se obtuvieron los tejidos a partir de la pechuga de pollo (*Gallus gallus*) de procedencias distintas.

En primer término, la muestra de tejido adiposo de *Gallus gallus*, tomada del Centro de Acopio de San Luis (CA001) tenía un color amarillento intenso y un olor característico a grasa fresco, no habían signos de enranciamiento, lo que garantiza que el producto es fresco y se mantuvo en las condiciones necesarias previa extracción del ARN.

La segunda muestra, proveniente del Mercado Modelo de Bellavista (MB001), tiene un color amarillo menos intenso, más parecido al color crema. Ninguna evidencia de enranciamiento previa extracción del ARN, lo que demuestra su frescura y que también se mantuvo en las condiciones adecuadas antes de su procesamiento.

4.2 Extracción de ARN

A continuación, en la figura 8, se muestra una imagen de la electroforesis realizada en gel de agarosa para la confirmación de la extracción de ARN total.



Figura 8. Gel de agarosa con banda de ARN de muestra CA001, visualizada en el transiluminador LED luz azul.

Para poder realizar la preparación de la librería y posterior secuenciamiento, el ARN extraído debe cumplir indispensablemente ciertos requisitos establecidos por el laboratorio que realizó el secuenciamiento, RIN entre 5 a 10 y una cantidad de ARN mayor a 1 µg. A continuación en la tabla N°7, se presentan los valores obtenidos de estas dos condiciones para las muestras CA001 y MB001.

Tabla N°7. Cantidad Total de ARN y valor de RIN de las muestras analizadas.

MUESTRA	CANTIDAD TOTAL DE ARN (µg)	RIN
CA001	1.480	6.6
MB001	1.504	5.7

4.3. Análisis Bioinformático

Los datos del secuenciamiento del transcriptoma se encuentran en 4 archivos (dos para cada muestra secuenciada) en formato FastQ (*.fq) correspondientes a las muestras de tejido adiposo CA001 y MB001, secuenciadas solamente una vez. Las condiciones y características de la secuenciación de ambos conjuntos de datos se describen en la tabla N°8.

Tabla N°8. Características de la tecnología utilizada para el secuenciamiento de las muestras CA001 y MB001.

INFORMACIÓN TÉCNICA DEL SECUENCIAMIENTO	
Tecnología de Secuenciamiento	RNA-sequencing / NGS (Next Generation Sequencing)
Plataforma de Secuenciamiento	Novaseq 6000 - Illumina
Kit de preparación	TruSeq Stranded Total RNA con Ribo Zero
Tipo de librería	Paired - end
Workflow	GenerateFASTQ
Cantidad de rendimiento	40 millones de lecturas por muestra (~4Gb).

4.3.1 Raw data

Los datos obtenidos según secuenciamiento fueron los expresados en la tabla N°9.

Tabla N°9. Resumen de la base de datos obtenida del secuenciamiento NGS.

ID de la muestra	Total de lecturas de bases (bp)	Total de lecturas	GC (%)	AT (%)	Q20 (%)	Q30 (%)
CA001	5,525,438,108	54,707,308	47.88	52.12	98.38	95.42
MB001	5,462,173,526	54,080,926	48.58	51.42	98.28	95.04

ID de la muestra: nombre de la muestra, **Total de lecturas de bases:** número total de bases secuenciadas, **Total de lecturas:** número total de lecturas. Para Illumina secuenciamiento tipo paired-end, este valor se refiere a la suma de las lecturas 1 y lectura 2 (forward y reverse), **GC (%):** contenido de GC, **AT (%):** contenido de AT, **Q20 (%):** proporción de bases que tienen un puntaje de calidad de phred de más de 20, **Q30 (%):** proporción de bases que tienen un puntaje de calidad de phred de más de 30.

4.3.2 Control de calidad de las lecturas

El análisis de control de calidad para las lecturas emparejadas fue hecho con el software FastQC, generándose resultados similares para cada uno de los ficheros en relación a los parámetros de calidad evaluados.

La figura 9 - A, muestra la distribución de los valores de calidad por base, la cual refiere que el 100% de las bases generadas en el secuenciamiento se encuentran por encima del valor de Phred 30, lo que indica una probabilidad de error del 0.1% por cada 100 bases, por lo tanto, las secuencias generadas por ambas muestras no necesitan de un filtrado posterior.

La figura 9 - B, muestra la distribución de puntuación de calidad en todas las secuencias. El gráfico señala claramente que el promedio de calidad de todas las secuencias según los valores de Phred es de 36, lo que concuerda con el anterior gráfico y brinda una garantía de calidad general de las secuencias. Finalmente la figura 9 - C, muestra el contenido de bases en la secuencia. En el gráfico se puede apreciar un ruido al iniciar el secuenciamiento de la lectura y líneas paralelas entre sí, esto es señal de normalidad en el proceso.

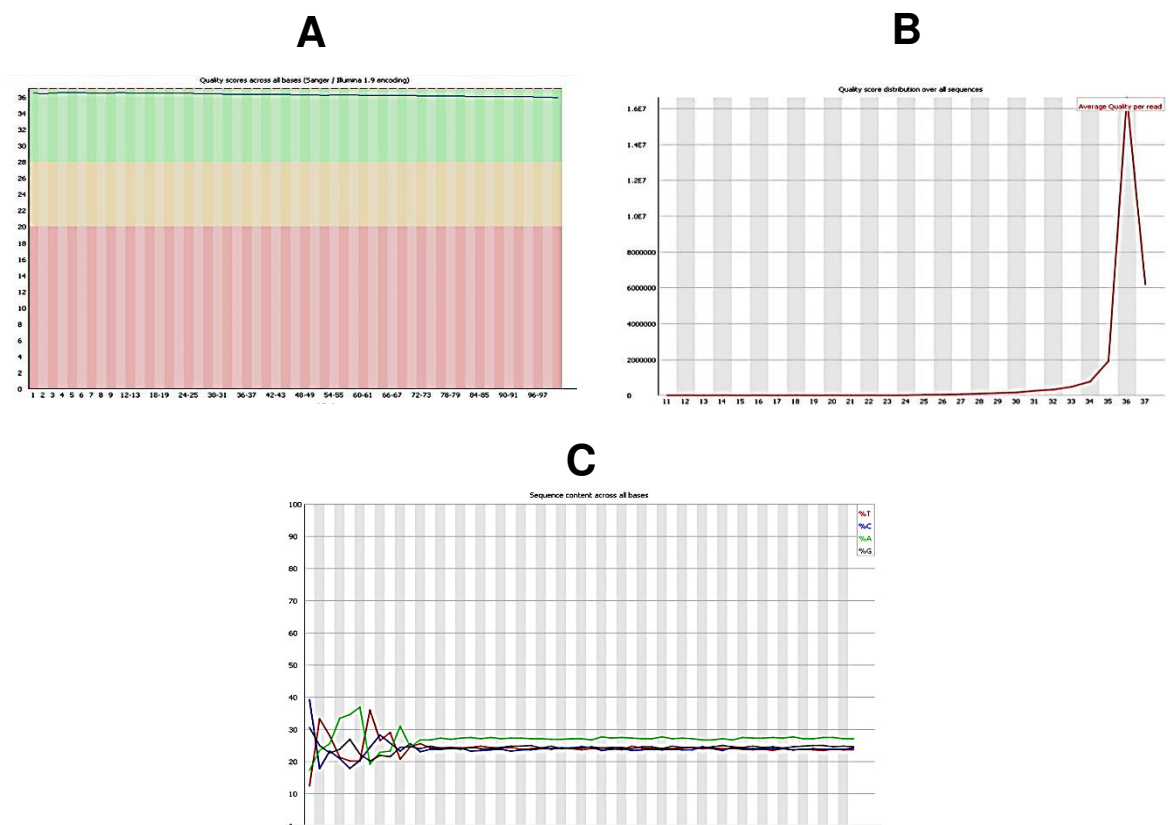


Figura 9. Gráficos del control de calidad de los transcritos secuenciados, generados por el programa FastQC.

4.3.3 Alineamiento de lecturas: Tophat

El software Tophat produce un número de archivos (4) en el directorio en el que se ejecutó, en nuestro estudio en la plataforma www.usegalaxy.org. La mayoría de estos archivos son intermedios o de transición, que se generan para su uso dentro de un pipeline o flujo de trabajo de análisis de RNA-seq. Los archivos de salida que se pueden obtener se muestran en imágenes y a continuación se describe su contenido.

En las figuras 10 y 11 se muestran los archivos `deletions.bed` e `insertions.bed`, los cuales reportan las exclusiones e inserciones reportadas por Tophat.

`Deletions.bed` – `chromLeft` (figura 10) se refiere a la primera base genómica de la eliminación, mientras `Insertions.bed` (figura 11) a la última base genómica antes de la inserción. Tal como lo indica cada una de las figuras, se puede apreciar la información del cromosoma, la descripción de las exclusiones e inserciones y las exclusiones e inserciones realizadas por Tophat.

En la figura 12, el archivo `junctions.bed`. En este archivo Tophat reporta las uniones. Cada unión consta de dos bloques BED conectados, donde cada bloque es tan largo como el saliente máximo de cualquier lectura que abarque la unión. Dentro del archivo encontramos la siguiente información:

- 1) Nombre del cromosoma (`chrom`)
- 2) La posición inicial (`Start`) y final (`End`) del rasgo en el cromosoma o scaffold.
- 3) El nombre (`name`) otorgado a una región (preferiblemente es único). Se usa "." si no hay un nombre asignado.
- 4) El número de alineamientos que abarcan la unión (`score`).
- 5) La hebra (`strand`), positiva (+) o negativa (-), para denotar filamento u orientación. Se usa "." si no hay orientación asignada.
- 6) La posición inicial en la que se dibuja el rasgo grueso (`thickStart`).
- 7) La posición final en la se dibuja el rasgo grueso (`thickEnd`).
- 8) El valor RGB (`Item RGB`) de la forma R, G, B (e.g. 255, 0, 0).
- 9) El número de bloques en la línea BED (`blockCount`).

- 10) Una lista separada por comas de los tamaños de bloque (BlockSizes). El número de elementos en esta lista debe corresponder a blockCount, en nuestro estudio corresponde a 2 bloques.
- 11) Una lista de bloques separados por comas (blockStarts). El primer es 0 y todas las posiciones de blockStart deben calcularse en relación con chromStart. El número de elementos en esta lista debe corresponder a blockCount.

Síntesis de los resultados obtenidos en los archivos de TopHat:

1) Resumen del alineamiento de la muestra CA001:

a) Lecturas desde la derecha:

- Entrada: 27353654
- Mapeadas: 23888699 (87% de lecturas de entrada)
- De las mapeadas: 689563 (2.9% tienen múltiples alineamientos, 4073 tienen >20).

b) Lecturas desde la izquierda:

- Entrada: 27353654
- Mapeadas: 24090451 (88.1% de lecturas de entrada).
- De las mapeadas: 726738 (3.0% tienen múltiples alineamientos, 4080 tienen >20).

Resultado global: 87.7% de la tasa general de lecturas mapeadas.

- Pares alineados: 22372984
- De los pares alineados: 628929 (2.8% tienen múltiples alineamientos), 185056 (0.8%) son alineamientos discordantes.

Resultado global: 81.1% de la tasa de alineamientos de pares concordantes.

2) Resumen del alineamiento de la muestra MB001:

a) Lecturas desde la derecha:

- Entrada: 27040463
- Mapeadas: 22217343 (82.2% de lecturas de entrada)
- De las mapeadas: 2000120 (9.0% tienen múltiples alineamientos, 18424 tienen >20).

b) Lecturas desde la izquierda:

- Entrada: 27040463
- Mapeadas: 22477644 (83.1% de lecturas de entrada).
- De las mapeadas: 2016556 (9.0% tienen múltiples alineamientos, 19424 tienen >20).

Resultado global: 82.6% de la tasa general de lecturas mapeadas.

- Pares alineados: 20979300
- De los pares alineados: 1900122 (9.1% tienen múltiples alineamientos), 104328 (0.5%) son alineamientos discordantes.

Resultado global: 77.2% de la tasa de alineamientos de pares concordantes

track name=deletions description="TopHat deletions"				
chr1	10859	10861	-	1
chr1	16163	16164	-	1
chr1	18470	18471	-	1
chr1	38813	38814	-	1
chr1	42973	42974	-	2
chr1	43144	43145	-	1
chr1	51394	51395	-	1
chr1	51937	51938	-	2
chr1	53506	53507	-	1
chr1	57274	57275	-	1
chr1	57378	57379	-	1
chr1	57453	57454	-	1
chr1	59192	59194	-	1
chr1	60484	60486	-	1
chr1	61248	61249	-	2
chr1	63607	63608	-	1
chr1	64137	64139	-	1
chr1	64920	64921	-	1
chr1	65001	65003	-	1
chr1	72858	72859	-	2
chr1	74259	74260	-	1

Figura 10. Deleciones en TopHat de muestra CA001

track name=insertions description="TopHat insertions"				
chr1	2371	2371	A	4
chr1	5824	5824	C	2
chr1	15240	15240	T	1
chr1	16283	16283	T	1
chr1	18590	18590	T	1
chr1	21975	21975	T	1
chr1	28289	28289	T	1
chr1	42959	42959	T	6
chr1	49147	49147	A	2
chr1	51393	51393	T	1
chr1	60484	60484	A	1
chr1	60485	60485	A	3
chr1	60566	60566	A	1
chr1	60985	60985	A	2
chr1	61234	61234	A	2
chr1	68302	68302	A	1
chr1	69575	69575	T	1
chr1	72513	72513	A	1
chr1	72584	72584	A	1
chr1	73838	73838	T	2
chr1	80910	80910	C	1

Figura 11. Inserciones en TopHat de muestra CA001

Chrom	Start	End	Name	Score	Strand	ThickStart	ThickEnd	ItemRGB	BlockCount	BlockSizes	BlockStarts
track name=junctions description="TopHat junctions"											
chr1	2187	9336	JUNC000000001	3	+	2187	9336	255,0,0	2	74,40	0,7109
chr1	2187	10845	JUNC000000002	3	+	2187	10845	255,0,0	2	74,40	0,8618
chr1	2371	9526	JUNC000000003	56	+	2371	9526	255,0,0	2	35,80	0,7075
chr1	2371	14755	JUNC000000004	54	+	2371	14755	255,0,0	2	78,58	0,12326
chr1	9420	14755	JUNC000000005	77	+	9420	14755	255,0,0	2	69,58	0,5277
chr1	14723	15032	JUNC000000006	2	+	14723	15032	255,0,0	2	76,25	0,284
chr1	15110	15490	JUNC000000007	85	+	15110	15490	255,0,0	2	79,99	0,281
chr1	41137	42508	JUNC000000008	30	+	41137	42508	255,0,0	2	92,99	0,1272
chr1	42414	43431	JUNC000000009	53	+	42414	43431	255,0,0	2	98,100	0,917
chr1	43384	44305	JUNC000000010	97	+	43384	44305	255,0,0	2	100,100	0,821
chr1	44309	45855	JUNC000000011	1	+	44309	45855	255,0,0	2	33,68	0,1478
chr1	44259	45869	JUNC000000012	9	+	44259	45869	255,0,0	2	87,82	0,1528
chr1	44259	45883	JUNC000000013	77	+	44259	45883	255,0,0	2	99,96	0,1528
chr1	45847	48532	JUNC000000014	65	+	45847	48532	255,0,0	2	99,100	0,2585
chr1	48455	49106	JUNC000000015	81	+	48455	49106	255,0,0	2	100,98	0,553
chr1	49022	51937	JUNC000000016	33	+	49022	51937	255,0,0	2	100,94	0,2821
chr1	52153	53538	JUNC000000017	40	+	52153	53538	255,0,0	2	99,80	0,1305
chr1	53456	55495	JUNC000000018	43	+	53456	55495	255,0,0	2	67,91	0,1948
chr1	55403	56623	JUNC000000019	51	+	55403	56623	255,0,0	2	92,76	0,1144
chr1	56547	57362	JUNC000000020	59	+	56547	57362	255,0,0	2	68,100	0,715
chr1	57497	57676	JUNC000000021	1	+	57497	57676	255,0,0	2	35,66	0,113
chr1	62206	63457	JUNC000000022	53	+	62206	63457	255,0,0	2	98,99	0,1152
chr1	65210	67369	JUNC000000023	46	+	65210	67369	255,0,0	2	99,99	0,2060
chr1	67350	68786	JUNC000000024	34	+	67350	68786	255,0,0	2	99,99	0,1337
chr1	68755	69796	JUNC000000025	28	+	68755	69796	255,0,0	2	98,100	0,941
chr1	69736	71088	JUNC000000026	36	+	69736	71088	255,0,0	2	92,96	0,1256
chr1	69754	71028	JUNC000000027	1	+	69754	71028	255,0,0	2	74,27	0,1247
chr1	69760	71378	JUNC000000028	3	+	69760	71378	255,0,0	2	68,95	0,1523
chr1	71053	71380	JUNC000000029	51	+	71053	71380	255,0,0	2	98,97	0,230
chr1	71282	72277	JUNC000000030	39	+	71282	72277	255,0,0	2	96,96	0,899
chr1	72217	73227	JUNC000000031	69	+	72217	73227	255,0,0	2	99,100	0,910

Figura 12. Uniones de empalme en TopHat (Muestra CA001)

QNAME	FLAG	RNAME	POS	MAPQ
@HD VN:1.0 SO:coordinate				
@SQ SN:chr1 LN:195276750				
@SQ SN:chr10 LN:19911089				
@SQ SN:chr10_AADN03010406_random LN:9054				
@SQ SN:chr10_AADN03010407_random LN:788				
@SQ SN:chr10_AADN03010408_random LN:1218				
@SQ SN:chr10_AADN03010409_random LN:1059				
@SQ SN:chr10_AADN03010410_random LN:9080				
@SQ SN:chr10_AADN03010415_random LN:8950				
@SQ SN:chr10_AADN03010416_random LN:11080				
@SQ SN:chr10_AADN03010417_random LN:11102				
@SQ SN:chr10_AADN03010418_random LN:8138				
@SQ SN:chr10_AADN03010419_random LN:1051				
@SQ SN:chr10_AADN03010420_random LN:8415				
@SQ SN:chr10_AADN03010422_random LN:817				
@SQ SN:chr10_AADN03010423_random LN:508				
@SQ SN:chr10_AADN03010424_random LN:564				
@SQ SN:chr10_AADN03010425_random LN:1440				
@SQ SN:chr10_AADN03010426_random LN:966				
@SQ SN:chr10_AADN03010427_random LN:1103				
@SQ SN:chr10_AADN03010428_random LN:633				
@SQ SN:chr10_AADN03010429_random LN:1791				
@SQ SN:chr10_AADN03010430_random LN:1640				
@SQ SN:chr10_JH375184_random LN:3009				
@SQ SN:chr10_JH375185_random LN:11223				
@SQ SN:chr11 LN:19401079				
@SQ SN:chr11_AADN03010432_random LN:15848				
@SQ SN:chr11_AADN03010433_random LN:2223				
@SQ SN:chr11_AADN03010434_random LN:1193				
@SQ SN:chr11_AADN03010435_random LN:1031				
@SQ SN:chr11_AADN03010436_random LN:2014				

Figura 13. Uniones Aceptadas en TopHat de muestra CA001

4.3.4 Ensamblaje de lecturas: Cufflinks

Procesa las lecturas aceptadas en formato BAM (Binary Alignment/Map Format), archivo generado por Tophat, de ambas muestras: CA001 y MB001.

Este software produce tres archivos de salida:

- 1) Como se aprecia en la figura 14, los niveles de expresión génica de ambas muestra en archivos genes.fpkm_tracking. Este archivo contiene los valores de la estimación de los niveles de expresión génica, individualmente, de cada una de las muestras, en el formato genérico FPKM (fragmentos por millón de kilobase) Tracking Format. La información importante que se puede conocer desde este archivo es el ID del gen en serie CUFF, el locus en el que se encuentra el gen, el límite inferior del intervalo de confianza del 95% de la abundancia de esta isoforma, es decir, $\text{límite inferior} = \text{FPKM} * (1.0 - \text{conf_lo})$ y el límite superior del intervalo de confianza del 95% de la abundancia de esta isoforma, es decir, $\text{límite superior} = \text{FPKM} * (1.0 + \text{conf_lo})$.
- 2) Seguidamente, en la figura 15, se observa la expresión de los transcritos, que posee la misma información que el archivo anterior, pero le agrega la longitud y cobertura de dicho transcrito en el locus señalado.
- 3) Finalmente los transcritos ensamblados que se aprecian en la figura 16, son dados a conocer a través del archivo transcripts.gtf. Este archivo GTF lleva por contenido las isoformas ensambladas por Cufflinks. Las primeras 7 columnas de este archivo son estándar GTF (Seqname, source, feature, start, end y score) y la última columna identifica los atributos, algunos de los cuales están también estandarizados ("gene_id", y "transcript_id"). Hay un registro GTF por fila, y cada registro representa una transcripción o un exón dentro de una transcripción.

1	2	3	4	5	6	7	8	9	10	11	12	13
tracking_id	class_code	nearest_ref_id	gene_id	gene_short_name	tss_id	locus	length	coverage	FPKM	FPKM_conf_lo	FPKM_conf_hi	FPKM_status
CUFF.1	-	-	CUFF.1	-	-	chr1:38792-40127	-	-	0.985657	0.60978	1.36153	OK
CUFF.3	-	-	CUFF.3	-	-	chr1:78246-78889	-	-	1.35797	0.650274	2.06566	OK
CUFF.5	-	-	CUFF.5	-	-	chr1:93911-94300	-	-	2.92858	1.30521	4.55194	OK
CUFF.6	-	-	CUFF.6	-	-	chr1:97909-98628	-	-	1.30974	0.674421	1.94505	OK
CUFF.7	-	-	CUFF.7	-	-	chr1:98695-99679	-	-	1.21688	0.716741	1.71701	OK
CUFF.4	-	-	CUFF.4	-	-	chr1:78942-80786	-	-	2.30854	1.07626	3.54081	OK
CUFF.2	-	-	CUFF.2	-	-	chr1:2266-15496	-	-	3.75187	2.2437	5.26004	OK
CUFF.9	-	-	CUFF.9	-	-	chr1:106354-110216	-	-	1.40776	0.977322	1.83819	OK
CUFF.11	-	-	CUFF.11	-	-	chr1:150415-152553	-	-	2.40394	1.96133	2.84654	OK
CUFF.8	-	-	CUFF.8	-	-	chr1:81336-87375	-	-	2.77295	1.95281	3.59308	OK
CUFF.10	-	-	CUFF.10	-	-	chr1:87499-89031	-	-	2.60686	1.93487	3.27886	OK
CUFF.12	-	-	CUFF.12	-	-	chr1:113451-120004	-	-	3.14447	2.71518	3.57375	OK
CUFF.13	-	-	CUFF.13	-	-	chr1:101337-106254	-	-	1.48685	1.16842	1.80528	OK
CUFF.14	-	-	CUFF.14	-	-	chr1:160600-163711	-	-	2.95216	2.52996	3.37436	OK
CUFF.15	-	-	CUFF.15	-	-	chr1:154032-160484	-	-	0.699668	0.517406	0.88193	OK
CUFF.16	-	-	CUFF.16	-	-	chr1:154032-160545	-	-	1.75566	1.50536	2.00595	OK
CUFF.17	-	-	CUFF.17	-	-	chr1:137173-140656	-	-	1.17934	0.760285	1.59839	OK
CUFF.18	-	-	CUFF.18	-	-	chr1:135023-140643	-	-	3.38766	3.01737	3.75796	OK
CUFF.19	-	-	CUFF.19	-	-	chr1:142145-143715	-	-	2.26089	1.73478	2.78699	OK
CUFF.20	-	-	CUFF.20	-	-	chr1:132574-132938	-	-	6.51732	3.87844	9.1562	OK
CUFF.21	-	-	CUFF.21	-	-	chr1:140768-142069	-	-	1.94324	1.40821	2.47826	OK
CUFF.22	-	-	CUFF.22	-	-	chr1:143848-150137	-	-	0.983832	0.770119	1.19755	OK
CUFF.23	-	-	CUFF.23	-	-	chr1:143848-150137	-	-	2.33376	2.02893	2.6386	OK
CUFF.24	-	-	CUFF.24	-	-	chr1:405187-405637	-	-	3.41828	1.92642	4.91014	OK
CUFF.26	-	-	CUFF.26	-	-	chr1:406409-409069	-	-	5.14683	4.56798	5.72568	OK
CUFF.27	-	-	CUFF.27	-	-	chr1:421593-422735	-	-	3.53591	2.62254	4.44929	OK
CUFF.28	-	-	CUFF.28	-	-	chr1:422856-424992	-	-	5.50624	4.82896	6.18351	OK
CUFF.25	-	-	CUFF.25	-	-	chr1:40183-76605	-	-	6.24464	0	26.3192	OK
CUFF.29	-	-	CUFF.29	-	-	chr1:409330-421491	-	-	4.51556	4.27072	4.7604	OK
CUFF.30	-	-	CUFF.30	-	-	chr1:528893-533092	-	-	5.17097	4.71634	5.6256	OK
CUFF.31	-	-	CUFF.31	-	-	chr1:618393-619026	-	-	7.69124	5.99136	9.39111	OK
CUFF.32	-	-	CUFF.32	-	-	chr1:666671-667199	-	-	1.35445	0.537688	2.17122	OK
CUFF.33	-	-	CUFF.33	-	-	chr1:673853-675701	-	-	1.24679	0.900758	1.59282	OK

Figura 14. Expresión Génica en Cufflinks de
Muestra CA001

1	2	3	4	5	6	7	8	9	10	11	12	13
tracking_id	class_code	nearest_ref_id	gene_id	gene_short_name	tss_id	locus	length	coverage	FPKM	FPKM_conf_lo	FPKM_conf_hi	FPKM_status
CUFF.1.1	-	-	CUFF.1	-	-	chr1:38792-40127	1335	4.26625	0.985657	0.60978	1.36153	OK
CUFF.3.1	-	-	CUFF.3	-	-	chr1:78246-78889	643	5.8165	1.35797	0.650274	2.06566	OK
CUFF.5.1	-	-	CUFF.5	-	-	chr1:93911-94300	389	11.5403	2.92858	1.30521	4.55194	OK
CUFF.6.1	-	-	CUFF.6	-	-	chr1:97909-98628	719	6.13791	1.30974	0.674421	1.94505	OK
CUFF.7.1	-	-	CUFF.7	-	-	chr1:98695-99679	984	5.8637	1.21688	0.716741	1.71701	OK
CUFF.4.1	-	-	CUFF.4	-	-	chr1:78942-80786	464	10.2835	2.30854	1.07626	3.54081	OK
CUFF.2.1	-	-	CUFF.2	-	-	chr1:2266-15496	572	5.36683	1.87593	0.81553	2.93634	OK
CUFF.2.2	-	-	CUFF.2	-	-	chr1:2266-15496	572	5.36683	1.87593	0.805389	2.94648	OK
CUFF.9.1	-	-	CUFF.9	-	-	chr1:106354-110216	1856	2.43378	0.564384	0.273337	0.855431	OK
CUFF.9.2	-	-	CUFF.9	-	-	chr1:107387-110216	1856	3.63685	0.843371	0.503224	1.18352	OK
CUFF.11.1	-	-	CUFF.11	-	-	chr1:150415-152553	2138	10.6282	2.40394	1.96133	2.84654	OK
CUFF.8.1	-	-	CUFF.8	-	-	chr1:81336-82597	679	4.74444	1.03667	0.411075	1.66227	OK
CUFF.8.2	-	-	CUFF.8	-	-	chr1:81336-87375	1981	1.82537	0.398847	0.145685	0.65201	OK
CUFF.8.3	-	-	CUFF.8	-	-	chr1:83812-87375	1679	3.86545	0.84461	0.481513	1.20771	OK
CUFF.8.4	-	-	CUFF.8	-	-	chr1:85995-87375	1008	2.25544	0.492819	0.0993763	0.886262	OK
CUFF.10.1	-	-	CUFF.10	-	-	chr1:87499-89031	1165	11.5333	2.60686	1.93487	3.27886	OK
CUFF.12.1	-	-	CUFF.12	-	-	chr1:113451-118936	3031	5.28771	1.09174	0.765061	1.41843	OK
CUFF.12.2	-	-	CUFF.12	-	-	chr1:113451-120004	3188	9.94207	2.05272	1.65346	2.45199	OK
CUFF.13.1	-	-	CUFF.13	-	-	chr1:101337-106254	3305	1.88626	0.382939	0.20523	0.560648	OK
CUFF.13.2	-	-	CUFF.13	-	-	chr1:101337-106254	3407	5.43757	1.10391	0.844974	1.36285	OK
CUFF.14.1	-	-	CUFF.14	-	-	chr1:160600-163711	3111	12.668	2.95216	2.52996	3.37436	OK
CUFF.15.1	-	-	CUFF.15	-	-	chr1:154032-160484	6231	2.95997	0.699668	0.517406	0.88193	OK
CUFF.16.1	-	-	CUFF.16	-	-	chr1:154032-160545	6513	7.42735	1.75566	1.50536	2.00595	OK
CUFF.17.1	-	-	CUFF.17	-	-	chr1:137173-140656	3060	4.85716	1.17934	0.83171	1.52696	OK
CUFF.18.1	-	-	CUFF.18	-	-	chr1:135023-140643	5460	13.9523	3.38766	3.01737	3.75796	OK
CUFF.19.1	-	-	CUFF.19	-	-	chr1:142145-143715	1570	9.3116	2.26089	1.73478	2.78699	OK
CUFF.20.1	-	-	CUFF.20	-	-	chr1:132574-132938	364	26.4121	6.51732	3.88919	9.14545	OK
CUFF.21.1	-	-	CUFF.21	-	-	chr1:140768-142069	1301	8.13839	1.94324	1.40821	2.47826	OK
CUFF.22.1	-	-	CUFF.22	-	-	chr1:143848-150137	6289	4.27582	0.983832	0.772991	1.19467	OK
CUFF.23.1	-	-	CUFF.23	-	-	chr1:143848-150137	6073	10.1427	2.33376	2.03136	2.63616	OK
CUFF.24.1	-	-	CUFF.24	-	-	chr1:405187-405637	450	14.6413	3.41828	1.92642	4.91014	OK
CUFF.26.1	-	-	CUFF.26	-	-	chr1:406409-409069	2660	23.1989	5.14683	4.56798	5.72568	OK
CUFF.27.1	-	-	CUFF.27	-	-	chr1:421593-422735	898	15.8292	3.53591	2.62254	4.44929	OK

Figura 15. Expresión de Transcritos en Cufflinks de Muestra CA001.

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
chr1	Cufflinks	transcript	38793	40127	1000	.	.	gene_id "CUFF.1"; transcript_id "CUFF.1.1"; FPKM "0.9856569885"; frac "1.000000"; conf_lo "0.609780"; conf_hi "1.361534"; cov "4.266249";
chr1	Cufflinks	exon	38793	40127	1000	.	.	gene_id "CUFF.1"; transcript_id "CUFF.1.1"; exon_number "1"; FPKM "0.9856569885"; frac "1.000000"; conf_lo "0.609780"; conf_hi "1.361534"; cov "4.266249";
chr1	Cufflinks	transcript	78247	78889	1000	.	.	gene_id "CUFF.3"; transcript_id "CUFF.3.1"; FPKM "1.3579667350"; frac "1.000000"; conf_lo "0.650274"; conf_hi "2.065659"; cov "5.816503";
chr1	Cufflinks	exon	78247	78889	1000	.	.	gene_id "CUFF.3"; transcript_id "CUFF.3.1"; exon_number "1"; FPKM "1.3579667350"; frac "1.000000"; conf_lo "0.650274"; conf_hi "2.065659"; cov "5.816503";
chr1	Cufflinks	transcript	93912	94300	1000	.	.	gene_id "CUFF.5"; transcript_id "CUFF.5.1"; FPKM "2.9285757386"; frac "1.000000"; conf_lo "1.305209"; conf_hi "4.551943"; cov "11.540300";
chr1	Cufflinks	exon	93912	94300	1000	.	.	gene_id "CUFF.5"; transcript_id "CUFF.5.1"; exon_number "1"; FPKM "2.9285757386"; frac "1.000000"; conf_lo "1.305209"; conf_hi "4.551943"; cov "11.540300";
chr1	Cufflinks	transcript	97910	98628	1000	.	.	gene_id "CUFF.6"; transcript_id "CUFF.6.1"; FPKM "1.3097363542"; frac "1.000000"; conf_lo "0.674421"; conf_hi "1.945052"; cov "6.137913";
chr1	Cufflinks	exon	97910	98628	1000	.	.	gene_id "CUFF.6"; transcript_id "CUFF.6.1"; exon_number "1"; FPKM "1.3097363542"; frac "1.000000"; conf_lo "0.674421"; conf_hi "1.945052"; cov "6.137913";
chr1	Cufflinks	transcript	98696	99679	1000	.	.	gene_id "CUFF.7"; transcript_id "CUFF.7.1"; FPKM "1.2168756355"; frac "1.000000"; conf_lo "0.716741"; conf_hi "1.717010"; cov "5.863697";
chr1	Cufflinks	exon	98696	99679	1000	.	.	gene_id "CUFF.7"; transcript_id "CUFF.7.1"; exon_number "1"; FPKM "1.2168756355"; frac "1.000000"; conf_lo "0.716741"; conf_hi "1.717010"; cov "5.863697";
chr1	Cufflinks	transcript	78943	80786	1000	+	.	gene_id "CUFF.4"; transcript_id "CUFF.4.1"; FPKM "2.3085351921"; frac "1.000000"; conf_lo "1.076258"; conf_hi "3.540812"; cov "10.283540";
chr1	Cufflinks	exon	78943	80786	1000	+	.	gene_id "CUFF.4"; transcript_id "CUFF.4.1"; exon_number "1"; FPKM "2.3085351921"; frac "1.000000"; conf_lo "1.076258"; conf_hi "3.540812"; cov "10.283540";
chr1	Cufflinks	transcript	80409	80482	1000	+	.	gene_id "CUFF.4"; transcript_id "CUFF.4.1"; exon_number "2"; FPKM "2.3085351921"; frac "1.000000"; conf_lo "1.076258"; conf_hi "3.540812"; cov "10.283540";
chr1	Cufflinks	exon	80409	80482	1000	+	.	gene_id "CUFF.4"; transcript_id "CUFF.4.1"; exon_number "2"; FPKM "2.3085351921"; frac "1.000000"; conf_lo "1.076258"; conf_hi "3.540812"; cov "10.283540";
chr1	Cufflinks	transcript	2267	15496	1000	+	.	gene_id "CUFF.2"; transcript_id "CUFF.2.1"; FPKM "1.8759327036"; frac "0.500000"; conf_lo "0.815530"; conf_hi "2.936335"; cov "5.366825";
chr1	Cufflinks	exon	2267	2449	1000	+	.	gene_id "CUFF.2"; transcript_id "CUFF.2.1"; exon_number "1"; FPKM "1.8759327036"; frac "0.500000"; conf_lo "0.815530"; conf_hi "2.936335"; cov "5.366825";
chr1	Cufflinks	exon	14698	14799	1000	+	.	gene_id "CUFF.2"; transcript_id "CUFF.2.1"; exon_number "2"; FPKM "1.8759327036"; frac "0.500000"; conf_lo "0.815530"; conf_hi "2.936335"; cov "5.366825";
chr1	Cufflinks	exon	15008	15189	1000	+	.	gene_id "CUFF.2"; transcript_id "CUFF.2.1"; exon_number "3"; FPKM "1.8759327036"; frac "0.500000"; conf_lo "0.815530"; conf_hi "2.936335"; cov "5.366825";
chr1	Cufflinks	exon	15392	15496	1000	+	.	gene_id "CUFF.2"; transcript_id "CUFF.2.1"; exon_number "4"; FPKM "1.8759327036"; frac "0.500000"; conf_lo "0.815530"; conf_hi "2.936335"; cov "5.366825";
chr1	Cufflinks	transcript	2267	15496	1000	+	.	gene_id "CUFF.2"; transcript_id "CUFF.2.2"; FPKM "1.8759327036"; frac "0.500000"; conf_lo "0.805389"; conf_hi "2.946477"; cov "5.366825";
chr1	Cufflinks	exon	2267	2406	1000	+	.	gene_id "CUFF.2"; transcript_id "CUFF.2.2"; exon_number "1"; FPKM "1.8759327036"; frac "0.500000"; conf_lo "0.805389"; conf_hi "2.946477"; cov "5.366825";
chr1	Cufflinks	exon	9447	9489	1000	+	.	gene_id "CUFF.2"; transcript_id "CUFF.2.2"; exon_number "2"; FPKM "1.8759327036"; frac "0.500000"; conf_lo "0.805389"; conf_hi "2.946477"; cov "5.366825";
chr1	Cufflinks	exon	14698	14799	1000	+	.	gene_id "CUFF.2"; transcript_id "CUFF.2.2"; exon_number "3"; FPKM "1.8759327036"; frac "0.500000"; conf_lo "0.805389"; conf_hi "2.946477"; cov "5.366825";
chr1	Cufflinks	exon	15008	15189	1000	+	.	gene_id "CUFF.2"; transcript_id "CUFF.2.2"; exon_number "4"; FPKM "1.8759327036"; frac "0.500000"; conf_lo "0.805389"; conf_hi "2.946477"; cov "5.366825";
chr1	Cufflinks	exon	15392	15496	1000	+	.	gene_id "CUFF.2"; transcript_id "CUFF.2.2"; exon_number "5"; FPKM "1.8759327036"; frac "0.500000"; conf_lo "0.805389"; conf_hi "2.946477"; cov "5.366825";
chr1	Cufflinks	transcript	106355	110216	669	+	.	gene_id "CUFF.9"; transcript_id "CUFF.9.1"; FPKM "0.5643842890"; frac "0.400911"; conf_lo "0.273337"; conf_hi "0.855431"; cov "2.433785";
chr1	Cufflinks	exon	106355	106478	669	+	.	gene_id "CUFF.9"; transcript_id "CUFF.9.1"; exon_number "1"; FPKM "0.5643842890"; frac "0.400911"; conf_lo "0.273337"; conf_hi "0.855431"; cov "2.433785";
chr1	Cufflinks	exon	106555	106769	669	+	.	gene_id "CUFF.9"; transcript_id "CUFF.9.1"; exon_number "2"; FPKM "0.5643842890"; frac "0.400911"; conf_lo "0.273337"; conf_hi "0.855431"; cov "2.433785";
chr1	Cufflinks	exon	107727	108695	669	+	.	gene_id "CUFF.9"; transcript_id "CUFF.9.1"; exon_number "3"; FPKM "0.5643842890"; frac "0.400911"; conf_lo "0.273337"; conf_hi "0.855431"; cov "2.433785";
chr1	Cufflinks	exon	109153	109250	669	+	.	gene_id "CUFF.9"; transcript_id "CUFF.9.1"; exon_number "4"; FPKM "0.5643842890"; frac "0.400911"; conf_lo "0.273337"; conf_hi "0.855431"; cov "2.433785";
chr1	Cufflinks	exon	109485	109568	669	+	.	gene_id "CUFF.9"; transcript_id "CUFF.9.1"; exon_number "5"; FPKM "0.5643842890"; frac "0.400911"; conf_lo "0.273337"; conf_hi "0.855431"; cov "2.433785";
chr1	Cufflinks	exon	109851	110216	669	+	.	gene_id "CUFF.9"; transcript_id "CUFF.9.1"; exon_number "6"; FPKM "0.5643842890"; frac "0.400911"; conf_lo "0.273337"; conf_hi "0.855431"; cov "2.433785";
chr1	Cufflinks	transcript	107388	110216	1000	+	.	gene_id "CUFF.9"; transcript_id "CUFF.9.2"; FPKM "0.8433710645"; frac "0.599089"; conf_lo "0.503224"; conf_hi "1.183518"; cov "3.636855";
chr1	Cufflinks	exon	107388	108695	1000	+	.	gene_id "CUFF.9"; transcript_id "CUFF.9.2"; exon_number "1"; FPKM "0.8433710645"; frac "0.599089"; conf_lo "0.503224"; conf_hi "1.183518"; cov "3.636855";
chr1	Cufflinks	exon	109153	109250	1000	+	.	gene_id "CUFF.9"; transcript_id "CUFF.9.2"; exon_number "2"; FPKM "0.8433710645"; frac "0.599089"; conf_lo "0.503224"; conf_hi "1.183518"; cov "3.636855";
chr1	Cufflinks	exon	109485	109568	1000	+	.	gene_id "CUFF.9"; transcript_id "CUFF.9.2"; exon_number "3"; FPKM "0.8433710645"; frac "0.599089"; conf_lo "0.503224"; conf_hi "1.183518"; cov "3.636855";
chr1	Cufflinks	exon	109851	110216	1000	+	.	gene_id "CUFF.9"; transcript_id "CUFF.9.2"; exon_number "4"; FPKM "0.8433710645"; frac "0.599089"; conf_lo "0.503224"; conf_hi "1.183518"; cov "3.636855";
chr1	Cufflinks	transcript	150416	152553	1000	.	.	gene_id "CUFF.11"; transcript_id "CUFF.11.1"; FPKM "2.4039350412"; frac "1.000000"; conf_lo "1.961335"; conf_hi "2.846535"; cov "10.628227";

Figura 16. Transcritos ensamblados en Cufflinks de Muestra CA001.

Tabla N°10. Cantidad de genes expresados en muestras CA001 y MB001 por Cufflinks

Muestra	Número de genes expresados
CA001	91993
MB001	74882

4.3.5 Expresión Diferencial Génica (DEG) en Cuffdiff

Cuffdiff calcula el FPKM de cada transcripción, transcripción primaria y gen en cada muestra. La transcripción primaria y el gen FPKM se calculan sumando los FPKM de las transcripciones en cada grupo de transcripción primaria o grupo de genes. Los resultados se envían en archivos de seguimiento FPKM en el formato descrito aquí. Hay cuatro archivos de seguimiento FPKM: isoforms.fpkm_tracking, genes.fpkm_tracking, cds.fpkm_tracking, tss_groups.fpkm_tracking.

A continuación, los resultados expuestos se han sintetizado a partir del archivo bisagra de esta investigación: genes.fpkm_tracking, que se basa en el FPKM sumado de las transcripciones que comparten cada ID del gen.

Tabla N°11. Evaluación de la expresión diferencial génica (p ajustado < 0.05) de CA001 con respecto a muestra MB001.

Genes	Cantidad
Sub-regulados (downregulated)	57
Sobre-regulados (upregulated)	6
Diferencial Infinito	71

Tabla N°12. Expresión génica diferencial (sub-regulados) más relevantes de la muestra CA001 con respecto a la muestra MB001 de tejido adiposo de *Gallus gallus*.

ID del Gen	Título	Locus	Log2(fold change)*	P valor ajustado (q valor)**	Participación en rutas metabólicas (KEGG)
PEPD	Peptidasa D	chr11:10450591-10473252	-3.97981	0.0333774	-
RPLP0	Subunidad P0 del tallo lateral de proteína ribosomal	chr15:9720482-9723290	-5.86048	0.0333774	Participa en la ruta metabólica del ribosoma.
CDK5RAP2	Subunidad regulatoria CDK5 asociado a la proteína 2	chr17:4877318-4880796	-6.33852	0.0333774	-
CUX1	Cortadora homeobox 1	chr19:4252923-4255792	-4.33825	0.0333774	-
OBSCN	Obscurina, calmodulina citoesquelética y RhoGEF que interactúa con la titina.	chr2:2185561-2206070	-4.84496	0.0333774	-
KLHL40	Miembro de la familia tipo-kelch 40	chr2:2094671-2125420	-4.46407	0.0333774	-
MRC1	Receptor de manosa tipo – C 1	chr2:19577891-19579946	-4.4976	0.0333774	Ruta metabólica de la formación del fagosoma.
ATXN1	Ataxin 1	chr2:62012768-62015056	-4.3683	0.0333774	-
NUP153	Nucleoporin 153	chr2:62044118-62110569	-3.80694	0.0333774	Dentro de la ruta de transporte de ARN, participa en la formación del complejo nuclear de poros.
ZNF236	Proteína dedo de zinc 236	chr2:90580940-90584551	-7.65757	0.0333774	-

NOL4	Proteína nucleonar 4	chr2:107228310-107231199	-3.48263	0.0333774	-
STAU2	Proteína Staufen de unión a ARN de doble cadena 2	chr2:117820358-117897191	-5.35795	0.0333774	-
TCEA2	Factor de alargamiento de la transcripción A2	chr20:9348884-9356226	-8.50317	0.0333774	-
CALD1	Caldesmon 1	chr1:62299210-62319999	-4.56764	0.0333774	Participa en la ruta metabólica de contracción del músculo liso vascular, directamente en la formación de actina.
COX7B	Citocromo c oxidasa subunidad 7B	chr4:12898196-12909390	-4.6369	0.0333774	El gen de la citocromo c oxidasa (COX), componente terminal de la cadena respiratoria mitocondrial, cataliza la transferencia de electrones del citocromo c reducido al oxígeno.
TET2	Tet metilcitosina dioxigenasa	chr4:38806819-38809843	-6.16458	0.0333774	-
BANK1	Proteína de andamio de células B con repeticiones de anquirina 1	chr4:60670289-60692675	-6.11292	0.0333774	Dentro de sus funciones como: supervivencia y muerte celular, desarrollo y función del sistema hematológico, respuesta inmune humoral.
LTK	Receptor de leucocitos tirosina quinasa	chr5:24962798-24964842	-5.01467	0.0333774	-
MEIS2	Meis homeobox 2	chr5:31546993-31551784	-6.5964	0.0333774	Está relacionado con el desarrollo del cerebro, corazón, ojos, cartílago y hematopoyesis.
PRKD1	Proteína quinasa D1	chr5:34118781-34283599	-3.57046	0.0333774	-
OVSTL	Ovostatin-like	chr1:76434208-76437372	-5.87822	0.0333774	-

EIF3F	Factor de iniciación de la traducción eucariótica 3 subunidad F	chr6:17046246-17064616	-3.5182	0.0333774	EI3F: dentro de la ruta de transporte del ARN como factor de iniciación de la traducción.
ZNF385B	Proteína de dedo de zinc 385B	chr7:14878920-14880523	-9.98605	0.0333774	-
ABL2	Protooncogén 2 ABL, tirosina quinasa no receptora	chr8:6609345-6609772	-4.17188	0.0333774	Participa en la ruta metabólica de ErbB. La familia ErbB de receptores tirosina quinasa (RTK) une a los ligandos del factor de crecimiento extracelular a las vías de señalización intracelular que regulan diversas respuestas biológicas.
FGGY	Dominio que contiene FGGY carbohidrato quinasa	chr8:26764297-26783622	-5.64444	0.0333774	-
SERINC5	Serina incorporadora 5	chrZ:21985257-21989427	-5.39709	0.0333774	-
THBS4	Trombospondina 4	chrZ:22016869-22032803	-5.97836	0.0333774	-

Tabla N°13. Expresión génica diferencial (sobre-regulados) de la muestra CA001 con respecto a la muestra MB001 de tejido adiposo de *Gallus gallus*.

ID del Gen	Título	Locus	Log2(fold change)*	p ajustado (q valor)**	Participación en rutas metabólicas (KEGG)
VPS50	Sindetin	chr2:23037756-23046081	4.99632	0.0333774	Actúa como componente del complejo EARP que está involucrado en el reciclaje endocítico. El complejo EARP se asocia con endosomas positivos para Rab4 y promueve el reciclaje del receptor de transferrina internalizado (TFRC) a la membrana plasmática.
TRNAV-AAC	Valina ARN de transferencia	chr2:89725704-89731481	4.62801	0.0333774	-
Gen no determinado	-	chr3:94589086-94596865	4.39706	0.0333774	-
Gen no determinado	-	chr3:94573480-94584011	4.16854	0.0333774	-
HLCS	Ligasa biotina-proteína	chr1:108275640-108299690	4.65006	0.0333774	Participa en el metabolismo de la Biotina, directamente relacionada con la formación de Holo (carboxilasa) que a su vez está ligada a la biosíntesis de ácidos grasos.
Gen no determinado	-	chr20:14256261-14283499	3.46018	0.0333774	-

Tabla N°14. Expresión génica diferencial más relevantes (con expresión igual a cero, Log2 = inf), de la muestra CA001 con respecto a la muestra MB001 de tejido adiposo de *Gallus gallus*.

ID del Gen	Título	Locus	Log2(fold change)*	P valor ajustado (q valor)**	Participación en rutas metabólicas (KEGG)
PLA2G10	Fosfolipasa A2	chr14:1242757-1243763	inf	0.0333774	Participa en el metabolismo del glicerofosfolípido, metabolismo de los lípidos, metabolismo del ácido araquidónico, metabolismo del ácido linoleico, metabolismo del ácido alfa linoleico y en la contracción vascular del músculo liso.
AXIN1	Axin1	chr14:12584496-12585204	inf	0.0333774	Participa en la vía de señalización Wnt.
CRHR2	Receptor de hormona liberadora de corticotropina 2	chr2:4274648-4275165	inf	0.0333774	Participa en la interacción neuroactiva del ligando – receptor.
DYNC1I1	Dineína citoplásmica 1 cadena intermedia 1	chr2:24029502-24030746	inf	0.0333774	Participa en la generación del fagosoma y en la ruta que sigue la infección con <i>Salmonella</i> .
ELMO1	Envolvimiento y motilidad celular 1	chr2:46520107-46521614	inf	0.0333774	-
PLCB1	Fosfolipasa C beta 1	chr3:14444624-14445303	inf	0.0333774	Participa en numerosas rutas, en las que destacan el metabolismo del fosfato inositol, la ruta de señalización del calcio, señalización adrenérgica en cardiomiocitos, ruta de señalización Wnt y en la ruta de señalización edad-rango en complicaciones

					diabéticas.
MAP3K1	Proteína quinasa quinasa - quinasa activada por mitógenos	chrZ:17424656-17425641	inf	0.0333774	Participa en la ruta de señalización MAPK, en la proteólisis mediada por Ubiquitina, ruta de señalización del receptor como RIG-I y en la ruta de señalización GnRH.
CDH13	Cadherina 13	chr11:16374427-16402968	inf	0.0333774	-
ITPR2	Receptor de inositol 1,4,5-trisfosfato tipo 2	chr1:67731518-67732480	inf	0.0333774	Participa en varias rutas, destacan: la ruta de señalización del calcio, meiosis de oocitos, senescencia celular, contracción vascular del músculo liso, ruta de señalización del apelin y ruta de señalización de lectina tipo - C
DKK2	Inhibidor de la vía de señalización dickkopf WNT 2	chr4:38240291-38241414	inf	0.0333774	Participa en la ruta de señalización de Wnt.
AKT3	Serina / treonina proteína quinasa no específica	chr3:35181754-35182907	inf	0.0333774	Participa en diversas rutas, entre las que destacan la apoptosis, senescencia celular, vía de señalización de la insulina, maduración de ovocitos mediada por progesterona, influenza A e infección por herpes simple.
GPR55	Receptor 55 acoplado a proteína G	chr9:5509447-5509607	inf	0.0333774	-

Tabla N°15. Expresión diferencial, no significativa, de genes representativos (lipogénicos e inmunológicos) en muestra CA001 con respecto a muestra MB001.

ID del Gen	Título	Locus	Log2(fold change)*	P valor ajustado (q valor)**	Participación en rutas metabólicas (KEGG)
FASN	Ácido graso sintasa	chr18:4923848-4936918	-0.30379	0.68765	Participa dentro del metabolismo de ácidos grasos, directamente, en la síntesis de ácido palmítico a partir del Malonil-CoA. En la biosíntesis de ácidos grasos participa activamente en la formación del butiril, hexanoil, ácido octanoico, ácido decanoico, entre otros. En la vía de señalización de la insulina participa de la lipogénesis.
PPARG	Receptor Gamma Activado por Proliferador de Peroxisoma	chr12:5236090-5238736	0.0506718	0.9471	Participa en la vía de señalización PPAR como un factor de transcripción de los adipocitos.
LPL	Lipoproteinlipasa	chrZ:54537102-54641352	1.08579	0.427	En el metabolismo de los glicerolípidos trabajada con los triacilglicelores, mientras que en la vía de señalización del PPAR cumple la función de transportador de ácidos grasos.
CCL1	C-C Ligando de quimioquinas 1	chr19:4932369-4941374	0.0150531	0.98515	Familia quimiocina, de la subfamilia CC, participan en la ruta de interacción del receptor citoquina – citoquina.
CCL15	Quimiocina	chr19:4932369-4941374	0.0150531	0.98515	Familia quimiocina, de la subfamilia CC, participan en la ruta de interacción del receptor citoquina – citoquina

* **Log2 (fold change):** el log(en base 2) de fold change de x/y. Fold change, es una medida que describe cuánto cambia una cantidad entre un original y una medición posterior. Se define como la relación entre las dos cantidades; para las cantidades A y B, entonces el cambio de B con respecto a A es B / A.

****p valor ajustado:** el valor de la tasa de descubrimientos falsos (false discovery rate) de p-valor en el test estadístico utilizando la corrección para múltiples evaluaciones de Benjamini-Hochberg.

V. DISCUSIÓN

La carne de pollo representa la fuente de proteína animal más consumida en el Perú, destacándose como la principal ciudad que consume dicho producto, la capital, Lima (18). Siendo esta carne la de mayor preferencia entre los consumidores peruanos, los diversos componentes de la cadena de producción buscan alternativas científicas para poder aprovechar mejor sus recursos y optimizar los costos de producción, muy aparte de brindar un producto de calidad (entiéndase inocuidad del producto y valor nutricional). Dentro de dicho marco, en la última década, el desarrollo de técnicas genómicas ha permitido el análisis, en general, de la variación de los niveles de expresión de genes, proteínas y metabolitos. Por tanto, las ómicas, principalmente la transcriptómica, proteómica y metabolómica, tienen un rol fundamental para poder producir un animal listo para el consumo en la menor cantidad de tiempo con las características preferidas por el consumidor (3).

La alimentación es un factor sumamente importante en el desarrollo del animal. La búsqueda de nuevas alternativas alimenticias que formen parte de la dieta del animal para poder obtener mayor cantidad de carcasa (músculo) y menor cantidad de grasa en el menor tiempo posible, representan un reto constante para la industria avícola. La raza, temperatura a la que están expuestos, higiene del corral, materiales de cama y la intensidad de la luz, son otros factores relevantes en el desarrollo de los pollos de engorde (138). El estudio del transcriptoma permitirá a partir de datos en masa obtener información de primera mano para poder tomar decisiones en la industria y saber a partir de qué componentes en la dieta y otros factores influyentes en el desarrollo del animal, ya mencionados, podrían incidir positivamente en la menor deposición de grasa sobre el músculo y ganancia de proteínas, de manera que la inversión de dinero sea menor, los tiempos de producción se acorten y el producto final tenga una calidad nutricional óptima.

En esta perspectiva, se planteó la investigación realizada con las muestras de mercado con un origen diferente, el cual permitió determinar en la expresión

diferencial de genes que se relacionan con rutas bioquímicas y metabólicas de ácidos grasos e inmunidad, que podrían tener impacto en la calidad del producto.

Como resultados, se analizan los resultados obtenidos por Cuffdiff (DEG) en relación a los objetivos e hipótesis planteados. En nuestro estudio MB001 es la muestra control y CA001 es la muestra experimental, de acuerdo a las condiciones del centro de venta que ya fueron mencionadas anteriormente. Dentro del archivo de Cuffdiff que nos muestra la expresión diferencial por medio del Log_2 de fold change, encontramos 3 resultados: los genes sobre-regulados que responden a un logaritmo positivo, los genes sub-regulados, que se relacionan con un logaritmo positivo y los genes que tienen un diferencial de expresión indeterminado, en los que para una de las muestras el nivel de expresión del gen es igual a cero (infinito). Se han considerado aquellos genes que tienen una expresión diferencial significativa determinado por el método de dispersión “ciego” aplicado por Cuffdiff (p corregido < 0.05) para poder debatir los resultados.

TopHat y Cufflinks son software que se utilizan normalmente dentro de un diseño de análisis de expresión génica diferencial en animales de interés para la producción (139,140) y han ganado una reputación bastante importante dentro de los investigadores que dedicados al estudio del transcriptoma. Las herramientas para el análisis diferencial varían; tenemos DEseq, edge R y, en nuestro flujo de trabajo, Cuffdiff, fue el elegido debido a las características que posee y beneficios que describe Trapnell *et al.*, 2012 (130).

En un estudio en el que también se trabajó el perfil de expresión génica del tejido adiposo del pollo de engorde (48), se identificaron 113 DEG sub-regulados y 220 DEG sobre-regulados, teniendo por nuestro lado 57 sub-regulados y 6 sobre-regulados, esta diferencia podría estar relacionada al número de réplicas biológicas, las condiciones de las muestras y el software para el análisis de expresión diferencial utilizado; en cuanto al número de réplicas, estas deben ser 6, y al menos 12 cuando es importante identificar genes diferencialmente expresados con un valor significativo (SDE, significantly differentially expressed). Se destaca a Cuffdiff como una herramienta básica en comparación con DESeq o Edge R, pero que para estudios con poca cantidad de muestras es muy útil (141).

En primer lugar, están los genes sub-regulados de CA001 en relación a MB001, descritos en la tabla N°12. Son 57 genes sub-regulados con una significancia ajustada menor a 0.05 (q), dentro de los cuales hay algunos con relevancia bioquímica para el animal. Se analizan los genes de mayor relevancia.

El primer gen de este grupo, RPLP0 (subunidad P0 del tallo lateral de proteína ribosomal), con un Log2 de -5.86048; su sobre-expresión está asociado con el metabolismo de proteínas, pero que en el caso de los pollos de engorde que han generado miopatía, enfermedad que trata de una inflamación muscular crónica acompañada de debilidad muscular, y está relacionada con dos defectos en la calidad de la carne de pollo conocidos como “franja blanca” (White striping, definición para la aparición de líneas blancas en el pecho) y “pecho leñoso” (woody breast, definición para una rigidez extrema de los filetes), los cuales generan una pérdida anual de 200 millones de dólares o más para la industria avícola estadounidense (142). En los resultados expresados en la tabla N°12, este gen está sub-regulado en la muestra CA001 en relación a la muestra MB001, lo cual no determina una mala calidad producto y haber tenido miopatía, ya que en el estudio mencionado, se analizaron una serie de 45 genes relacionados. Otro gen perteneciente a este grupo de genes candidatos, es el CDH13 (cadherin 13), el cual se encuentra inactivo en la muestra MB001, pero muy activo en la muestra CA001 (tabla N°14), lo que indicaría que se trata de un metabolismo normal de proteínas al no tener expresión diferencial de otros genes inmersos en el metabolismo de proteínas.

Otro de los genes sub-regulados de importancia, es el MRC 1 (receptor de manosa tipo – C 1), con un Log2 de -4.4976, del cual su sobre-regulación, permite la eliminación de las glucoproteínas que dañan el tejido del huésped de la circulación en mamíferos (143,144) y participa en la ruta de formación de fagosomas, de los cuales se sabe son parte de un mecanismo de defensa (respuesta inmune) ante microorganismos patógenos en mamíferos, pero del cual no se han reportado resultados de los efectos de una sub-regulación en pollos de engorde. Por otra parte, daría soporte a los estudios que afirman que el tejido graso de *Gallus gallus* es un órgano inmunológico y puede alterar la expresión de genes relacionados a esta función de acuerdo a la edad y por suplementos

nutricionales como la arginina (48,145). Hasta ahora, es poco el conocimiento que se posee sobre sus funciones biológicas, pero se puede inferir que la sobre-expresión de MRC1 aviar puede contribuir de forma similar a la eliminación de la inflamación (146).

El siguiente gen con alta relevancia es el denominado MEIS2 (Homeobox Meis 2), con un Log2 de -6.5964, se le ha asociado con el desarrollo del cerebro, ojos, cartílago, corazón y hematopoyesis (147). Gheyas *et al.*, 2015 (148), indica que en su estudio detectaron 17 ventanas adyacentes como firmas de selección putativas (GGA5: 30, 490,001–30, 690,000) que albergan el gen MEIS2, que se unieron en una enorme región de 200 kb, siendo una fuerte región de selección. Un gen de gran importancia que debería ser estudiado a profundidad. Su sub-regulación se debería al acelerado desarrollo de la muestra CA001 y su tamaño menor en comparación a la muestra MB001.

Luego, también tenemos el gen ZNF385B, con un Log2 de -7.65757, el cual ha sido estudiado en pollos Arkansas Regressor (AR) y Arkansas Progressor (AP), siendo estos importantes modelos animales para el estudio de bases moleculares resistencia a enfermedades y susceptibilidad. El gene en mención tiene por ubicación el núcleo celular y está relacionado con un grupo de “genes candidatos”, a la muerte y supervivencia celular, al desarrollo del sistema hematológico y a la respuesta inmune humoral. Este grupo de genes candidatos, incluyendo ZNF385B, están ligados dentro de una red de trabajo (network) a la ubiquitina C (UBC), que tiene como funciones: degradación de las proteínas, reparación del ADN, regulación del ciclo celular, modificación de la kinasa, endocitosis y regulación de otras rutas de señalización (149).

El otro grupo de genes, son los sobre-regulados, descritos en la tabla N°13. Para este grupo se identificaron 6 locus que indicaban sobre-regulación según el valor del fold change. Al realizarse la búsqueda en el *Genome Data Viewer* (visualizador de la información del genoma) de tres genes no determinados (no se tiene codificación), se realizó la verificación en el buscador mencionado, corroborando la falta de información.

Dentro del grupo de genes sobre-regulados, hay uno con marcada relevancia: HLCS, con un Log2 de 4.65006, que codifica la proteína holocarboxilasa sintetasa. Esta participa en el metabolismo de la biotina y está involucrada en muchas funciones celulares, como la producción y catabolismo de proteínas, grasas y carbohidratos como parte del mismo metabolismo en mención. Como se puede ver en la ruta descrita en el KEGG (150), esta holocarboxilasa, está vinculada directamente a la biosíntesis de ácidos grasos. Por el lado de nuestros resultados, su sobreexpresión en la muestra CA001 en relación a la muestra MB001 puede tener diferentes posibles causas.

En primer punto, la biotina funciona como una coenzima en la hidrólisis de la síntesis de ácidos grasos y aminoácidos a través de reacciones de carboxilación (151). Es complicado que se presente deficiencia, debido a que las bacterias naturales de la flora intestinal del animal están preparadas para producir cantidades suficientes de esta vitamina (152). Su deficiencia puede ser el resultado de que el animal ha recibido suplementos, en alta cantidad, de proteínas antagonistas con efectos negativos: escaso crecimiento, debilidad de las patas, aparición de costras, adelgazamiento de la piel de las patas, infertilidad, entre otros problemas reproductivos (153). En su tesis, Castilla, 2018 (154), hizo una evaluación de un grupo de pollos al que se le añadió a su dieta un suplemento nutricional líquido que contenía 2 mg de biotina. No evidenció diferencias significativas en los grupos analizados en relación al control que consumía su dieta normal, sin el suplemento, lo que indicaría que normalmente el pollo produce la cantidad necesaria de esta vitamina.

Un dato importante, que podría estar muy relacionado con los resultados obtenidos en esta investigación con relación a la expresión de la biotina, es lo que señalan en su investigación Memon *et al.*, 2014 (155), que indica que un elevado nivel de biotina está asociado a un incremento de consumo de alimentos por parte del animal debido a que la biotina es una coenzima participante en el metabolismo de carbohidratos y está relacionada en la conversión de proteína y carbohidrato en grasas, lo que resulta en un incremento del apetito y posterior ingesta de alimentos.

Otro grupo dentro los resultados de expresión diferencial, son los genes que se mantienen inactivos en la muestra control. MB001 (transcritos = 0) y los que sí han mostrado actividad en la muestra experimental CA001 (número de transcritos sobre 0). Son 1807 genes los que muestran estos niveles de transcripción en ambas muestras, por lo tanto el logaritmo en base 2 de fold change será infinito ("inf", se lee en la tabla de resultados). Encontramos 71 genes que muestran una diferencia de transcripción significativa dentro de este grupo (p ajustado < 0.05). Hay, como en los demás grupos, genes que juegan un rol fundamental con respecto a la codificación de una proteína con una función fisiológica de relevancia o en participación en alguna ruta metabólica.

Dentro de los genes más relevantes dentro de este grupo (tabla N°14) tenemos el gen PLA2G10 que codifica la proteína Grupo 10 fosfolipasa secretora A2. Esta enzima sintetizada por este gen participa en diferentes rutas metabólicas: metabolismo de glicerofosfolípidos, metabolismo del lípido éter, metabolismo del ácido araquidónico, metabolismo del ácido linoleico, metabolismo del ácido alfa linoleico y en la contracción del músculo liso vascular (150). En su tesis, Lea, 2013 (156), reporta que la ruta metabólica del ácido linoleico se vio afectada por la ingesta de 800 g/t de NCF (fracción natural de carbohidratos) comparado con la muestra control. Esta fosfolipasa, en específica, reconoce el enlace acilo en la posición de los fosfolípidos sn-2 (fosfolipasa A2) y libera ácido araquidónico y fosfolípidos oxidados para generar los productos pro-inflamatorios que están implicados en la disfunción endotelial, la inflamación de la placa, y la formación del núcleo necrótico de la placa, y se postula que está vinculada a las modificaciones oxidativas de las LDL y el desarrollo de respuestas inflamatorias en la íntima arterial. En ese marco, la fosfolipasa A2, grupo 10 (PLA2G10) **mostró un valor de 0 (cero) para la expresión de este gen en la muestra MB001.**

Luego, tenemos el gen AXIN1, el cual al ser estudiado en líneas experimentales consanguíneas de aves, identificó múltiples locis de tratamiento cuantitativo (QTL) asociados con la resistencia a la colonización, ya sea por *Salmonella* como *Campylobacter*, junto con los genes NRAMP1 y TRAP1 (157,158). Se puede también apreciar su participación en la ruta de señalización WNT, en la que es la

codificación de una proteína de andamiaje que controla los niveles de β -catenina. También presenta inactividad (expresión igual a 0) para la muestra MB001.

Otro de los genes que están inactivos en la muestra MB001 es el CRHR2 (Receptor de hormona liberadora de corticotropina 2) que codifica la proteína del mismo nombre. Participa en la ruta de interacción neuroactiva ligando-receptor. Como indica en su estudio Li *et al.*, 2017 (159), este gen se presenta sub-regulado en condiciones de estrés por frío, al igual que Kang & Kuenzuel, 2017 (160), que señala lo mismo, bajo los efectos de un estrés crónico.

El gen DYNC1I1 codifica la proteína Dineína citoplasmática 1 - cadena intermedia 1, que participa en las rutas de generación del fagosoma e infección de la *Salmonella*. No se reportan estudios con relación a estas rutas metabólicas, pero Gu *et al.*, 2011 (161), reportan que un polimorfismo dentro de este gen, localizado en 23.9 Mb (GGA2), está asociado a la ganancia de peso corporal.

El gen PLCB1 es otro de los cuales también se encuentra activo en la muestra CA001 e inactivo en la muestra MB001 (expresión génica igual a 0). Participa en 11 rutas metabólicas: metabolismo del inositol fosfato, ruta de señalización del calcio, sistema de señalización de la fosfatidilinositol, señalización adrenérgica en cardiomiocitos, contracción del músculo liso vascular, ruta de señalización WNT, ruta de señalización del Apelin, vía de señalización del receptor similar a NOD, melanogénesis, vía de señalización de GnRH, vía de señalización del AGE-RAGE en las complicaciones diabéticas (150).

El gen DKK2 (Inhibidor de la vía de señalización de Dickkopf WNT 2) codifica la proteína del mismo nombre que participa en la ruta de señalización WNT. Kong *et al.*, 2017 (162), estudiaron la expresión diferencial génica en el músculo de la pechuga de pollos de engorde PeM moderno (rápido crecimiento, alta eficiencia de alimentación) en comparación con la raza BPR no seleccionada (crecimiento lento, baja eficacia de la alimentación). Dentro de los genes sub-regulados en este estudio se encuentra el DKK2 y se incluye dentro del grupo de genes que secretan proteínas. También, tiene un rol en la diferenciación de osteoblastos en matrices mineralizadas. La inactivación en la muestra MB001 y su actividad en la muestra CA001, no permiten determinar el nivel de expresión diferencial. Se

podría sugerir un estudio más profundo sobre este grupo de genes asociados con el crecimiento del músculo y disminución de tejido adiposo en los pollos de engorde y la relación, por defecto, con la eficiencia de la dieta administrada.

En la última tabla (Tabla N°15) podemos apreciar genes de relevancia pero que no muestran una expresión diferencial significativa entre las muestras. Tenemos los genes lipogénicos como FASN que codifica la proteína del mismo nombre y que es único que está sub-regulado según el valor en la muestra CA001 con relación a la muestra MB001. Los otros dos genes relacionados a la lipogénesis (LPL y PPAR) se muestran sobre-regulados según los valores de fold change, pero en todos los casos sin lograr diferencias significativas. Por ejemplo, Xing y Jiang, 2011 (47), realizaron un estudio en el que evaluaron los efectos sobre los genes lipogénicos, entre ellos los mencionados, al añadir betaína como suplemento a la dieta del pollo de engorde. Según sus resultados y los reportados por otros autores (163), la betaína reduce considerablemente la deposición de grasa vía reducción de la lipogénesis, ergo, la sub-regulación del gen FASN. Con respecto a PPARG, se expresó en una ligera mayor proporción para el grupo experimental, al parecer esta sobreexpresión estaría ligada a la edad del animal. Finalmente el gen LPL también redujo su expresión debido al suplemento de betaína. Cabe indicar que los valores siguen siendo positivos y son muy cercanos al valor de 1, cuando hubo reducción en la expresión de los genes mencionados en el estudio de Xing y Jiang, 2011 (47).

Los otros dos genes dentro de este grupo que no poseen una expresión diferencial significativa son CCL1, CCL15. En el estudio de Hausmann *et al.*, 2014 (48), obtuvieron una sobreexpresión en relación a la edad del animal. El animal de 6 semanas presentaba una sobreexpresión del gen en relación a su par de 3 semanas de vida. Considerando sus resultados no solo de los genes mencionados que no tienen una expresión diferencial significativa en la relación muestra MB001 con CA001, el estudio de Hasuman *et al.*, 2014 (48), concluye que los genes expresados a partir del ARN de tejido graso de pollos de engorde, harían necesario considerarlo como un órgano inmunológico. En dicho estudio se compararon los niveles de expresión génica en relación a la edad del animal, factor fundamental para poder determinar niveles de expresión de diferentes

genes, debido a que se es “una fotografía del momento” y las necesidades son diferentes según la etapa de vida.

En los que corresponde a otros factores críticos para la obtención de los niveles de expresión génica a partir del transcriptoma, está el proceder de las muestras. La muestra CA001 proviene de un Centro de Acopio que no cumple con las condiciones de higiene y salubridad necesarias para el expendio de alimentos frescos, en este caso, pollos, incluso el animal puede haber pasado por cuadros de estrés debido a que se encuentran en jaulas de plástico hacinados. La muestra MB001 fue tomada en un mercado que cumple con las condiciones de higiene y saneamiento, tanto instalaciones como manipuladores. A partir de esto, se puede inferir que los métodos de crianza son distintos y que la expresión de genes relacionados al estrés o a respuestas inmunes para cada una de estas muestras seleccionadas, es diferente. Incluso, las dietas administradas pueden haber sido distintas para poder optimizar los tiempos de crecimiento y desarrollo.

Es importante resaltar que el valor del RIN del ARN que será utilizado para el secuenciamiento de las muestras debe ser de preferencia mayor a 7, aunque es válido un valor desde 5; y la cantidad de ARN debe ser mayor a 1µg, según indicaciones del secuenciador y del laboratorio en el que se realizó el secuenciamiento, ya que estos valores indican una estabilidad adecuada para el flujo de trabajo. Un valor de 10 indica un ARN intacto, un valor de 5 indica ARN parcialmente degradado, mientras que un valor menor a 3 es evidencia de ARN bastante degradado (164). Las muestras de ARN utilizadas para esta investigación presentan un valor de RIN de 6.6 para CA001 y 5.7 para MB001, lo cual indicaría que la integridad del ARN está dentro de parámetros aceptables. La parcial degradación del ARN podría deberse al tiempo que tomó la muestra en llegar hasta las instalaciones del laboratorio donde sería secuenciada o también a la susceptibilidad del ARN, ya que es propenso a degradarse a causa de las ARNasas del medio ambiente.

El tipo de librería seleccionado, paired end, ofrece mayor información para el análisis. Permite secuenciar ambos extremos (5' y 3') de un fragmento y obtener datos de secuencias alineables de alta calidad. Este tipo de secuenciamiento facilita la detección de reordenamientos genómicos y elementos de secuencias

repetitivas, así como transcritos de *nov*o (165). Lo corrobora Sengupta *et al.*, 2011(166), quienes indican que el enfoque paired – end permite ubicar uniones de empalme (splice junctions), y, nuevamente, su capacidad de ensamblar transcritos de *nov*o.

La plataforma tecnológica utilizada para el secuenciamiento fue el NovaSeq 6000, que ofrece una salida de data mayor a 6 Tb y 20 B en menos de dos días, por tanto las 40 millones de lecturas por muestra (4Gb) están dentro del rango establecido de lo que esta plataforma tecnológica de Illumina puede generar.

El control de calidad de los 4 archivos o files Fastq (2 por cada muestra) fue realizado por el programa FastQC, el cual genera una serie de gráficos que garantizan la calidad de las lecturas o si es que hay que realizar algún filtrado para que el análisis posterior pueda ser efectivo y preciso. El programa FastQC es una herramienta bioinformática frecuentemente utilizada por los investigadores debido a su facilidad de manejo y la información generada es la necesaria para poder concluir si se puede dar el siguiente paso o hay que filtrar ciertas lecturas (167). Según Andrews, 2010 (101), los resultados dentro de los gráficos de calidad de la secuencia por base, puntajes de calidad por base, contenido de la secuencia por bases y distribución de la longitud de la secuencia, representan valores confiables para poder realizar el análisis posterior con los software especializados para el RNA-seq.

La plataforma a partir de la cual se accedió a los software utilizados para el desarrollo de este estudio, fue www.usegalaxy.org; es una plataforma de acceso libre en internet, en la cual cualquier usuario interesado puede trabajar en los software disponibles, solamente cargando sus archivos. Esta plataforma dentro de cada software permite editar los parámetros de análisis, como se realizó en TopHat, Cufflinks y Cuffdiff para este estudio, uno puede editarlos de acuerdo a lo que necesita y requiere. Hay sistemas operativos, como windows a través de MobaXtern por ejemplo o Linux que permiten un análisis más profundo, en donde el programador puede establecer los parámetros de acuerdo al lenguaje operativo que maneja como Perl, Python o Java. Es ahí donde el filtrado, alineamiento, ensamble y otras funciones se pueden determinar los parámetros de mejor forma, además permite el análisis de una mayor cantidad de datos y muestras. La

plataforma utilizada para este estudio, a pesar de ser solamente 2 muestras, consume largo tiempo y demanda de una amplia capacidad computacional. Son diferencias que podrían brindar una mayor calidad de resultados para futuros estudios de contar con estaciones computacionales con mayor capacidad, velocidad y resolución.

El primer software utilizado fue TopHat. Este se encarga del alineamiento de las lecturas en relación a un genoma de referencia. Cabe precisar que el genoma de referencia utilizado es el Galgal 4.0 (*Gallus_gallus*-4.0), que está disponible en la plataforma web www.usegalaxy.org desde noviembre del 2011 y es el más utilizado hasta ahora, ya que recién en octubre del 2018 se ha liberado el ensamblado de Galgal 5.0 (*Gallus_gallus*- 5.0), que contiene mayor información en cuanto a genes, 4679 anotados, tanto transcritos de codificación de proteína como no transcritos (42) y 34 cromosomas identificados contra los 31 identificados en la versión Galgal 4.0 (168). Esto definitivamente podría interferir en el alineamiento y seguida pérdida de información de los transcritos debido a la menor cantidad de genes y cromosomas identificados por el genoma de referencia Galgal 4.0. Cabe indicar que en el archivo de expresión diferencial de Cuffdiff solamente se identifican genes hasta el cromosoma 28.

No existen muchos estudios de análisis del transcriptoma del tejido adiposo de *Gallus gallus* como tal, como indica Hausmman *et al.*, 2014 (48), hacen falta estudios que eluciden el rol que tienen los genes expresados en el tejido adiposo con el crecimiento y sistema inmune del animal.

El alineamiento de las lecturas secuenciadas en relación a un genoma de referencia es uno de los pasos fundamentales en los flujos de trabajo para ensayos de análisis de secuenciamiento de alto rendimiento (114). TopHat utiliza otro programa, Bowtie, como un impulsor de alineamiento, para posteriormente rompe las lectura que Bowtie no puede alinear por sí solo en piezas más pequeñas, denominadas segmentos. Estas piezas son las que se alinean con el genoma. Las alineaciones se pueden usar para cuantificar la expresión de genes y transcripciones, porque el número de lecturas producidas por una transcripción es proporcional a su abundancia. Por ejemplo, Yi *et al.*, 2016 (169), realizaron el perfil del transcriptoma del músculo de la pechuga de *Gallus gallus* en el que

utilizaron como software de alineamiento a TopHat y también como genoma de referencia el *Gallus gallus* 4.0. TopHat, para este estudio genera una tasa de mapeo del 70% en promedio, mientras que para nuestra investigación obtenemos una tasa de mapeo del 84% en promedio. Esto debido a la mayor cantidad de lecturas mapeadas (entre 40 y 60 millones), mientras que para nuestro estudio las lecturas mapeadas oscilan entre 27 y 28 millones. Esta variación significativa tiene su origen en los raw reads generados; mientras que para nuestro estudio estuvieron bordeando los 55 millones, las muestras de Yi *et al.*, 2016 (169), generaron entre 70 y 88 millones de raw reads. Factores como la muestra (músculo) y la plataforma de secuenciamiento son fundamentales para explicar las diferencias. En el caso de la investigación que se menciona, se utilizó el HiSeq 2000, los métodos de extracción de ARN y el kit de preparación de la librería, fueron los mismos que se usaron en nuestro estudio.

El siguiente análisis le correspondió al software Cufflinks, este es el encargado de generarlos archivos de expresión génica y prepara así los archivos que servirán para los software que son de su misma familia como el Cuffmerge o el Cuffdiff, que ha sido utilizado en este estudio para el análisis de la expresión diferencial. Lo importante que representa Cufflinks como software, es lo que señala Conesa *et al.*, 2016 (68), que este software ha sido diseñado especialmente para tomar ventaja de las lecturas paired – end, y que puede utilizar información de un formato GTF para identificar transcritos expresados, o puede inferir transcritos de *novo* desde un dataset mapeado sin genoma de referencia. Es el complemento perfecto para el genoma generado por un mapeador como TopHat y suelen encontrarse juntos en los flujos de análisis RNA –Seq. Dentro de los resultados que emite, está la expresión génica de cada una de las muestras que serán comparadas para la expresión diferencial génica por Cuffdiff. Para la muestra CA001 Cufflinks detecta 91995 genes, mientras que para MB001, son 74884 genes. La diferencia y en cuanto a estos niveles de expresión pueden ser causa de la integridad (menor por parte de MB001) del ARN (RIN).

Para concluir el análisis bioinformático, se aplica el software Cuffdiff, de la familia *Cuff* y la cual ha tenido por varios años la preferencia de los investigadores, conocido como “The Tuxedo pipeline”.

Como señala Trapanell *et al.*, 2012 (130), Cuffdiff se encarga de calcular la expresión en dos o más muestras y evalúa la significancia estadística de cada modificación observada en la expresión entre las muestras. Cuffdiff genera numerosos archivos donde se puede analizar la expresión diferencial de las muestras. Estos archivos contienen un conjunto de datos estadísticos que permiten identificar la sub-regulación o sobre-regulación de los genes mapeados y alineados. Uno de esos valores es el fold change (o en español, cambios en la superposición de genes) en logaritmo base 2 o binario. El uso del logaritmo binario, es útil cuando la información a calcular es la contraria; como por ejemplo, cuántas posiciones binarias y se necesitarán si se requiere codificar x datos, entre otros. Luego el P valor (ambos valores, el inicial y el corregido para múltiples evaluaciones como valor de q). Finalmente, se encuentran los atributos asignados a los genes y transcritos, tales como su codificación en Cuffdiff y la ubicación en el genoma (locus).

Finalmente, a pesar de que se ha trabajado con algunas limitantes como la susceptibilidad del ARN a las condiciones en las que se mantuvo hasta su secuenciación y la plataforma virtual www.usegalaxy.org, que permite un análisis con ciertas restricciones, se pudo determinar que existen genes expresados diferencialmente en la muestra CA001 con respecto a la muestra MB001, de lo que se puede inferir que posiblemente factores externos en el periodo de crecimiento de ambas muestras, como el nivel de estrés o la dieta determinada tendrían impacto en la síntesis de ácidos grasos y disposición final de tejido adiposo en la pechuga del animal.

VI. CONCLUSIONES

- Existen diferencias en el transcriptoma de tejido graso de *Gallus gallus* de las muestras analizadas procedentes de diferentes centros de venta, posiblemente asociados a factores como la crianza y la fisiología del animal.
- Se determinó la presencia de 57 genes sub-regulados (downregulated), destacando los genes MEIS2 que está relacionado con el desarrollo del cerebro, corazón, ojos, cartílago y hematopoyesis, y MRC1 que participa en la ruta de la formación del fagosoma.
- Se determinó la presencia de 6 genes sobre-regulados (upregulated), destacando el gen HLCS (holo-carboxilasa) que participa en la ruta del metabolismo de la biotina.
- Existen 71 genes inactivos en la muestra MB001, los cuales se encuentran activos en la muestra CA001, destacando el gen Fosfolipasa A2, propaga la inflamación mediante la producción de precursores del ácido araquidónico.
- Se identificó que genes previamente reportados como participantes activos dentro del metabolismo de los lípidos, como FASN, LPL y PPARG no tienen una expresión diferencial significativa en la muestra de tejido adiposo de CA001 con respecto a la muestra MB001.

VII. RECOMENDACIONES

Según los resultados obtenidos se recomienda lo siguiente:

- Realiza un estudio similar con mayor cantidad de muestras y repeticiones biológicas por cada una para tener una base de datos sólida.
- Hacer un estudio comparativo con diferentes software para cada una de los procesos realizados en esta investigación, como el mapeo, alineamiento y, sobre todo, la expresión diferencial génica.
- Realizar un modelo de estudio que permita controlar la dieta del animal, el tiempo de crecimiento y la temperatura del medio ambiente al que está sometido para determinar qué factores y variables tienen mayor relación con una expresión génica diferencial.
- Realizar el mismo modelo de estudio en otros órganos como el hígado o el músculo de donde proviene la proteína.
- Realizar estudios comparativos con otras ómicas vinculadas como la epigenómica, proteómica y metabolómica.

VIII. REFERENCIAS BIBLIOGRÁFICAS

1. Asociación Peruana de Avicultura (APA). No Title [Internet]. 2015 [cited 2019 Jan 25]. Available from: <http://www.apa.org.pe/html/nuestros-servicios-estadistica.php>.
2. Ministerio de Agricultura y Riego [MINAGRI]. Boletín Estadístico Mensual de la Producción y Comercialización de Productos Avícolas, Setiembre - Noviembre 2018. Lima, Perú; 2018 Dec.
3. Zampiga M, Flees J, Meluzzi A, Dridi S, Sirri F. Application of omics technologies for a deeper insight into quali-quantitative production traits in broiler chickens: A review. J Anim Sci Biotechnol [Internet]. 2018 Sep 10;9:61.
4. Frith MC, Pheasant M, Mattick JS. Genomics: The amazing complexity of the human transcriptome. Eur J Hum Genet [Internet]. 2005 Jun 22;13:894.
5. Henson J, Tischler G, Ning Z. Next-generation sequencing and large genome assemblies. Pharmacogenomics [Internet]. 2012 Jun;13(8):901–15.
6. Zhang Y, Li D, Han R, Wang Y, Li G, Liu X, et al. Transcriptome analysis of the pectoral muscles of local chickens and commercial broilers using Ribo-Zero ribonucleic acid sequencing. PLoS One [Internet]. 2017 Sep 1;12(9):e0184115–e0184115.
7. Boone C, Mourot J, Grégoire F, Remacle C. The adipose conversion process: regulation by extracellular and intracellular factors. Reprod Nutr Dev [Internet]. 2000 Jul;40(4):325–58.
8. Fève B. Adipogenesis: cellular and molecular aspects. Best Pract Res Clin Endocrinol Metab [Internet]. 2005;19(4):483–99.
9. Cui HX, Zheng MQ, Liu RR, Zhao GP, Chen JL, Wen J. Liver dominant expression of fatty acid synthase (FAS) gene in two chicken breeds during intramuscular-fat development. Mol Biol Rep [Internet]. 2012;39(4):3479–84.

10. Li Q, Li N, Hu X, Li J, Du Z, Chen L, et al. Genome-wide mapping of DNA methylation in chicken. PLoS One [Internet]. 2011 May 5;6(5):e19428–e19428.
11. Banco Central de Reserva del Perú [BCRP]. Panorama actual y proyecciones macroeconómicas 2017-2018 (Reporte de inflación). Lima, Perú; 2017.
12. Ministerio de Agricultura y Riego [MINAGRI]. Boletín estadístico mensual de la producción y comercialización avícola (Año 2). Lima, Perú; 2016 Dec.
13. Instituto Nacional de Estadística e Informática [INEI]. Producción nacional: Diciembre 2016. Lima, Perú; 2017 Feb.
14. Havenstein GB, Ferket PR, Qureshi MA. Carcass composition and yield of 1957 versus 2001 broilers when fed representative 1957 and 2001 broiler diets. Poult Sci [Internet]. 2003 Oct 1;82(10):1509–18.
15. Havenstein GB, Ferket PR, Qureshi MA. Growth, livability, and feed conversion of 1957 versus 2001 broilers when fed representative 1957 and 2001 broiler diets. Poult Sci [Internet]. 2003 Oct 1;82(10):1500–8.
16. Avicultura.com. Crecimiento constante del consumo de carne de pollo en Perú. 2016.
17. Arellano R. El pollo de cada día. El Comercio [Internet]. 2016 Oct 24; Available from: <http://elcomercio.pe/opinion/rincon-del-autor/pollo-dia-rolando-arellano-273971>
18. Bustamante Mejía C, Gutiérrez Villacorta E, Malla Alcalde J, Palacios Pando R, Revilla Medina F. Planeamiento Estratégico para el Subsector del Pollo de Engorde en el Perú. Universidad Católica del Perú; 2018.
19. What Are Broiler Chickens? [Internet]. 2015 [cited 1BC Mar 5]. Available from: <https://www.chickencheck.in/farm-to-table/what-are-broiler-chickens/>
20. Proyectos Peruanos. Crianza de pollos de carne [Internet]. 2017. Available from: http://proyectosperuanos.com/pollos_de_carne_cobb_ross/

21. Ministerio de Agricultura y Riego [MINAGRI]. Boletín estadístico de producción agrícola, pecuaria, y avícola. Lima, Perú; 2016.
22. Ministerio de Agricultura y Riego [MINAGRI]. Boletín estadístico mensual de la producción y comercialización avícola (Año 3). Lima, Perú; 2017.
23. Salas E. Aves [Internet]. Available from: https://tarwi.lamolina.edu.pe/~emellisho/zootecnia_archivos/produccion_aves.pdf
24. Ministerio de Agricultura y Riego [MINAGRI]. Diagnóstico de crianzas priorizadas para el plan ganadero 2017-2021. Lima, Perú; 2017.
25. El Sitio Avícola. Tendencias avícolas mundiales 2016: América representa el 44% de la producción mundial de pollo. [Internet]. 2016 [cited 2019 Jan 24]. Available from: <http://www.elsitioavicola.com/articles/2866/tendencias-avacolas-mundiales-2016-amarica-representa-el-44-por-ciento-de-la-produccion-mundial-de-pollo/>
26. Boehringer Ingelheim Perú. Investigación de mercado interna: Unidad de Salud Animal. Lima, Perú; 2017.
27. Rojas I. Ganadera Santa Elena se fusiona con Avinka asistida por Payet, Rey [Noticias] [Internet]. Lex Latin. 2017 [cited 1BC Jan 27]. Available from: <http://lexlatin.com/noticia/ganadera-santa-elena-se-fusiona-avinka-asesoriapayet-rey/>
28. Perú y Brasil son los países que más pollo consumen en América Latina [Internet]. Proexpansión. 2014 [cited 1BC Jan 20]. Available from: <http://proexpansion.com/es/articles/765-peru-y-brasil-son-los-paises-que-maspollo-consumen-en-america-latina>
29. Soriano J. Comunicación Personal. 2017.
30. Oficina de Estudios Económicos y Estadística. Estadística de Comercialización de Aves Vivas en Lima Metropolitana. Lima, Perú; 2018.
31. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, et al.

- Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* [Internet]. 2004 Dec 9;432:695.
32. Darwin C. *The Variation of Animals and Plants Under Domestication*. London; 1868.
 33. Fumihito A, Miyake T, Sumit S, Takadat M, Ohno S. One subspecies of the red junglefowl (*Gallus gallus gallus*) suffices as the matriarchic ancestor of all domestic breeds. *Proc Natl Acad Sci USA*. 1994;91:12505–9.
 34. FAO. *The State of the World's Animal Genetic Resources for Food and Agriculture*. FAO Rome; 2007.
 35. Fumihito A, Miyake T, Takada M, Shingu R, Endo T, Gojobori T, et al. Monophyletic origin and unique dispersal patterns of domestic fowls. *Proc Natl Acad Sci U S A* [Internet]. 1996 Jun 25;93(13):6792–5.
 36. Schmid M, Smith J, Burt DW, Aken BL, Antin PB, Archibald AL, et al. Third Report on Chicken Genes and Chromosomes 2015. *Cytogenet Genome Res* [Internet]. 2015/07/14. 2015;145(2):78–179.
 37. Rubin C-J, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* [Internet]. 2010 Mar 10;464:587.
 38. Fu W, Lee WR, Abasht B. Detection of genomic signatures of recent selection in commercial broiler chickens. *BMC Genet* [Internet]. 2016 Aug 26;17(1):122.
 39. ITIS. *Gallus gallus* (Linnaeus, 1758) [Internet]. [cited 1BC Mar 11]. Available from:
https://www.itis.gov/servlet/SingleRpt/SingleRpt?search_topic=TSN&search_value=176086#null
 40. NCBI. *Gallus gallus* (chicken) [Internet]. [cited 1BC Mar 11]. Available from:
<https://www.ncbi.nlm.nih.gov/genome/?term=Gallus+gallus>
 41. Burt DW. The chicken genome. *Genome Dyn*. 2006;2(2):123–37.

42. Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, et al. A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. *G3(Bethesda)* [Internet]. 2017;7(1):109–17.
43. Trayhurn P. Endocrine and signalling role of adipose tissue: new perspectives on fat. *Acta Physiol Scand* [Internet]. 2005 Aug 1;184(4):285–93.
44. Schäffler A, Schölmerich J. Innate immunity and adipose tissue biology. *Trends Immunol*. 2010;31(iv):228–35.
45. Bourneuf E, Hérault F, Chicault C, Carré W, Assaf S, Monnier A, et al. Microarray analysis of differential gene expression in the liver of lean and fat chickens. *Gene* [Internet]. 2006;372:162–70.
46. Wang H-B, Li H, Wang Q-G, Zhang X-Y, Wang S-Z, Wang Y-X, et al. Profiling of chicken adipose tissue gene expression by genome array. *BMC Genomics* [Internet]. 2007 Jun 27;8:193.
47. Xing J, Kang L, Jiang Y. Effect of dietary betaine supplementation on lipogenesis gene expression and CpG methylation of lipoprotein lipase gene in broilers. *Mol Biol Rep*. 2011;1975–81.
48. Hausman GJ, Barb CR, Fairchild BD, Gamble J, Lee-Rutherford L. Gene expression profiling in adipose tissue from growing broiler chickens. *Adipocyte* [Internet]. 2014 Dec 10;3(4):297–303.
49. Vailati-Riboni M, Palombo V, Loor JJ. What Are Omics Sciences? In: Ametaj BN, editor. *Periparturient Diseases of Dairy Cows*. 2017. p. 1–7.
50. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol* [Internet]. 2017 May 5;18(1):83.
51. Braconi D, Bernardini G, Millucci L, Santucci A. Foodomics for human health: current status and perspectives. *Expert Rev Proteomics* [Internet]. 2018 Feb 1;15(2):153–64.
52. García-Cañas V, Simó C, Herrero M, Ibáñez E, Cifuentes A. Present and

- Future Challenges in Food Analysis: Foodomics. *Anal Chem* [Internet]. 2012 Dec 4;84(23):10150–9.
53. Cowburn G, Stockley L. Consumer understanding and use of nutrition labelling: a systematic review. *Public Health Nutr* [Internet]. 2007/01/02. 2005;8(1):21–8.
 54. Gaso-Sokac D, Kovač S, Josic D. Application of proteomics in food technology and food biotechnology: Process development, quality control and product safety. Vol. 48, *Food Technology and Biotechnology*. 2010. 284-295 p.
 55. Gaso-Sokac D, Kovač S, Josic D. Use of Proteomic Methodology in Optimization of Processing and Quality Control of Food of Animal Origin. Vol. 49, *Food Technology and Biotechnology*. 2011. 397-412 p.
 56. Andjelković U, Šrajer Gajdošik M, Gašo-Sokač D, Martinović T, Josić D. Foodomics and Food Safety: Where We Are. *Food Technol Biotechnol* [Internet]. 2017 Sep;55(3):290–307.
 57. Transcriptomics [Internet]. Available from: www.nature.com/subjects/transcriptomics
 58. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57–63.
 59. Clark TA, Sugnet CW, Ares M. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science*. 2002;296(5569):907–10.
 60. Okoniewski MJ, Miller CJ. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*. 2006;7(Mm):1–14.
 61. Royce TE, Rozowsky JS, Gerstein MB. Toward a universal microarray: Prediction of gene expression through nearest-neighbor probe sequence identification. *Nucleic Acids Res*. 2007;35(15):1–10.

62. Gerhard D, Wagner L, Feingold E, Al. E. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res.* 2004;14(10B):2121–7.
63. Boguski C, Tolstochev C, Bassett Jr D. Gene discovery in dbEST. *Am Assoc Adv Sci.* 1994;265(5181):1993–4.
64. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* [Internet]. 2000 Jun 1;18:630.
65. Peiffer JA, Kaushik S, Sakai H, Arteaga-Vazquez M, Sanchez-Leon N, Ghazal H, et al. A spatial dissection of the Arabidopsis floral transcriptome by MPSS. *BMC Plant Biol* [Internet]. 2008;8(1):43.
66. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* [Internet]. 2008/05/01. 2008 Jun 6;320(5881):1344–9.
67. Morin RD, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh TJ, et al. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques.* 2008;45(1):81–94.
68. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol* [Internet]. 2016 Jan 26;17:13.
69. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* [Internet]. 2010/12/30. 2011 Feb;12(2):87–98.
70. Tachibana C. Transcriptomics today: Microarrays, RNA-seq, and more. *Science* (80-) [Internet]. 2015 Jul 31;349(6247):544 LP-546.
71. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Rep* [Internet]. 2012;2(3):666–73.
72. Holt RA, Jones SJM. The new paradigm of flow cell sequencing10.1101/gr.073262.107. *Genome Res* [Internet].

2008;18(6):839–46.

73. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* [Internet]. 2008 May 2;133(3):523–36.
74. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* [Internet]. 2008 Sep;18(9):1509–17.
75. Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* [Internet]. 2008 May 30;5:613.
76. Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS. SNP discovery via 454 transcriptome sequencing. *Plant J* [Internet]. 2007 Sep;51(5):910–8.
77. Emrich SJ, Barbazuk WB, Li L, Schnable PS. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res* [Internet]. 2007 Jan;17(1):69–73.
78. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, et al. Nucleotide sequence of bacteriophage ϕ X174 DNA. *Nature* [Internet]. 1977;265(5596):687–95.
79. Collins FS, Morgan M, Patrinos A. The Human Genome Project: Lessons from Large-Scale Biology. *Science* [Internet]. 2003 Apr 11;300(5617):286 LP-290.
80. Lander ES. Initial impact of the sequencing of the human genome. *Nature* [Internet]. 2011 Feb 9;470:187.
81. Besser J, Carleton HA, Gerner-Smidt P, Lindsey RL, Trees E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin Microbiol Infect* [Internet]. 2017/10/23. 2018 Apr;24(4):335–41.
82. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*

[Internet]. 2008 Oct 9;26:1135.

83. May GD, Thudi M, Varshney RK, Jackson SA, Li Y. Current state-of-art of sequencing technologies for plant genomics research. *Brief Funct Genomics* [Internet]. 2012 Feb 7;11(1):3–11.
84. Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. *Genome Res* [Internet]. 2010 Sep;20(9):1165–73.
85. Metzker ML. Sequencing technologies — the next generation. *Nat Rev Genet* [Internet]. 2009 Dec 8;11:31.
86. Kulski J. Next-generation sequencing-An overview of the history, tools, and “Omic” applications, next generation sequencing-advances, applications and challenges. *InTech*. 2016.
87. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. *J Appl Genet* [Internet]. 2011/06/23. 2011 Nov;52(4):413–35.
88. Kchouk M, Gibrat JF, Elloumi M. Generations of Sequencing Technologies: From First to Next Generation. *Biol Med* [Internet]. 2017;09(03).
89. Dark MJ. Whole-genome sequencing in bacteriology: state of the art. *Infect Drug Resist* [Internet]. 2013 Oct 8;6:115–23.
90. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! *Genome Biol* [Internet]. 2011 Aug 25;12(8):125.
91. Kasarskis A, Turner S, Schadt EE. A window into third-generation sequencing. *Hum Mol Genet* [Internet]. 2010 Sep 21;19(R2):R227–40.
92. Vincent AT, Derome N, Boyle B, Culley AI, Charette SJ. Next-generation sequencing (NGS) in the microbiological world: How to make the most of your money. *J Microbiol Methods* [Internet]. 2017;138:60–71.
93. Chaitankar V, Karakulah G, Ratnapriya R, Giuste FO, Brooks J, Swaroop A. Next generation sequencing technology and genomewide data analysis:

Perspectives for retinal research [Internet]. Progress in Retinal and Eye Research. Elsevier Ltd; 2016.

94. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature [Internet]. 2009 Feb 1;458:223.
95. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, et al. The developmental transcriptome of *Drosophila melanogaster*. Nature [Internet]. 2010/12/22. 2011 Mar 24;471(7339):473–9.
96. Yaragatti M, Basilico C, Dailey L. Identification of active transcriptional regulatory modules by the functional assay of DNA from nucleosome-free regions. Genome Res [Internet]. 2008 Jun;18(6):930–8.
97. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, et al. High-Resolution Profiling of Histone Methylations in the Human Genome. Cell [Internet]. 2007 May 18;129(4):823–37.
98. Deurenberg RH, Bathoorn E, Chlebowicz MA, Couto N, Ferdous M, García-Cobos S, et al. Application of next generation sequencing in clinical microbiology and infection prevention. J Biotechnol [Internet]. 2017;243:16–24.
99. Kumar R, Ichihashi Y, Kimura S, Chitwood DH, Headland LR, Peng J, et al. A High-Throughput Method for Illumina RNA-Seq Library Preparation. Front Plant Sci [Internet]. 2012 Aug 28;3:202.
100. Conesa A, García-Alcalde F, Dopazo J, Carbonell J, Okonechnikov K, Cruz LM, et al. Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics [Internet]. 2012 Aug 22;28(20):2678–9.
101. Andrews S. FASTQC. A quality control tool for high throughput sequence data. [Internet]. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
102. Dai M, Thompson RC, Maher C, Contreras-Galindo R, Kaplan MH, Markovitz DM, et al. NGSQC: cross-platform quality analysis pipeline for

- deep sequencing data. BMC Genomics [Internet]. 2010 Dec 2;11 Suppl 4(Suppl 4):S7–S7.
103. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics [Internet]. 2012/10/25. 2013 Jan;29(1):15–21.
 104. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol [Internet]. 2013;14(4):R36.
 105. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics [Internet]. 2009/03/16. 2009 May 1;25(9):1105–11.
 106. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics [Internet]. 2010/02/10. 2010 Apr 1;26(7):873–81.
 107. Jean G, Kahles A, Sreedharan VT, De Bona F, Rätsch G. RNA-seq read alignments with PALMapper. Current Protocols in Bioinformatics. 2010. 1-37 p.
 108. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res [Internet]. 2010/08/27. 2010 Oct;38(18):e178–e178.
 109. Marco-Sola S, Sammeth M, Guigó R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods [Internet]. 2012 Oct 28;9:1185.
 110. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol [Internet]. 2014/04/20. 2014 May;32(5):462–4.
 111. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics [Internet]. 2014/09/25. 2015 Jan

15;31(2):166–9.

112. Smyth GK, Shi W, Liao Y. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* [Internet]. 2013 Nov 13;30(7):923–30.
113. Bioinformatics UG. Frequently Asked Questions: Data File Formats. [Internet]. Available from: <https://genome.ucsc.edu/FAQ/FAQformat.html#format4>
114. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* [Internet]. 2010/05/02. 2010 May;28(5):511–5.
115. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* [Internet]. 2011 Aug 4;12:323.
116. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods* [Internet]. 2012/11/18. 2013 Jan;10(1):71–3.
117. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* [Internet]. 2016 Apr 4;34:525.
118. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* [Internet]. 2010/03/02. 2010;11(3):R25–R25.
119. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* [Internet]. 2010;11(10):R106.
120. Nueda MJ, Tarazona S, Conesa A. Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics* [Internet]. 2014/06/03. 2014 Sep 15;30(18):2598–602.
121. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical

- methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics [Internet]. 2010;11(1):94.
122. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics [Internet]. 2009/11/11. 2010 Jan 1;26(1):139–40.
 123. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol [Internet]. 2014/12/05. 2014;15(12):550.
 124. Sonesson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics [Internet]. 2013;14(1):91.
 125. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol [Internet]. 2013/09/10. 2013;14(9):R95–R95.
 126. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform [Internet]. 2013/12/02. 2015 Jan;16(1):59–70.
 127. Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics [Internet]. 2012;13(1):484.
 128. Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlén M, et al. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*. Nucleic Acids Res [Internet]. 2012/09/08. 2012 Nov;40(20):10084–97.
 129. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol [Internet].

2014;15(2):R29.

130. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* [Internet]. 2012 Mar 1;7(3):562–78.
131. Servicio Nacional de Sanidad Agraria [SENASA]. Reglamento del Sistema Sanitario Avícola. Perú; 2007 p. 7–18.
132. Bacci G. Raw Sequence Data and Quality Control BT - Bacterial Pangenomics: Methods and Protocols. In: Mengoni A, Galardini M, Fondi M, editors. New York, NY: Springer New York; 2015. p. 137–49.
133. Ewing B, Hillier L, Wendl MC, Green P. Base-Calling of Automated Sequencer Traces Using Phred . I . Accuracy Assessment. *Genome Res*. 1998;8(3):175–85.
134. Walther D, Bartha G, Morris M. Basecalling with LifeTrace. *Genome Res* [Internet]. 2001 May;11(5):875–88.
135. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, et al. Whole-genome sequencing and variant discovery in *C. elegans*. *Nat Methods* [Internet]. 2008 Jan 20;5:183.
136. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* [Internet]. 2008 Nov;18(11):1851–8.
137. Trapnell C. Cufflinks: Transcriptome assembly and differential expression analysis for RNA-Seq [Internet]. 2017 [cited 1BC Apr 22]. Available from: <http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/index.html>
138. Glatz P, Rym R. Poultry housing and management in developing countries. *Poultry development review*. Food Agric Organ. 2013;1–5.
139. Dado-Senn B, Skibieli AL, Fabris TF, Zhang Y, Dahl GE, Peñagaricano F, et al. RNA-Seq reveals novel genes and pathways involved in bovine mammary involution during the dry period and under environmental heat

- stress. *Sci Rep* [Internet]. 2018;8(1):11096.
140. Armstrong E, Iriarte A, Nicolini P, De Los Santos J, Ithurralde J, Bielli A, et al. Comparison of transcriptomic landscapes of different lamb muscles using RNA-Seq. *PLoS One* [Internet]. 2018 Jul 24;13(7):e0200732.
 141. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* [Internet]. 2016 Jun;22(6):839–51.
 142. Hargis BM, Kuttappan VA, Owens CM. White striping and woody breast myopathies in the modern poultry industry: a review. *Poult Sci* [Internet]. 2016 Jul 21;95(11):2724–33.
 143. Lee SJ, Evers S, Roeder D, Parlow AF, Risteli J, Risteli L, et al. Mannose Receptor-Mediated Regulation of Serum Glycoprotein Homeostasis. *Science* (80-) [Internet]. 2002 Mar 8;295(5561):1898 LP-1901.
 144. Gazi U, Martinez-Pomares L. Influence of the mannose receptor in host immune responses. *Immunobiology* [Internet]. 2009;214(7):554–61.
 145. Tan J, Liu S, Guo Y, Applegate TJ, Eicher SD. Dietary l-arginine supplementation attenuates lipopolysaccharide-induced inflammatory response in broiler chickens. *Br J Nutr* [Internet]. 2014/01/13. 2014;111(8):1394–404.
 146. Kraaij MD, van Dijk A, Haagsman HP. CATH-2 and LL-37 increase mannose receptor expression, antigen presentation and the endocytic capacity of chicken mononuclear phagocytes. *Mol Immunol* [Internet]. 2017;90:118–25.
 147. Machon O, Masek J, Machonova O, Krauss S, Kozmik Z. Meis2 is essential for cranial and cardiac neural crest development. *BMC Dev Biol* [Internet]. 2015 Nov 6;15:40.
 148. Gheyas AA, Boschiero C, Eory L, Ralph H, Kuo R, Woolliams JA, et al. Functional classification of 15 million SNPs detected from diverse chicken populations. *DNA Res* [Internet]. 2015/04/29. 2015 Jun;22(3):205–17.

149. Khatri B, Hayden AM, Anthony NB, Kong BC. Whole Genome Resequencing of Arkansas Progressor and Regressor Line Chickens to Identify SNPs Associated with Tumor Regression. *Genes (Basel)* [Internet]. 2018 Oct 19;9(10):512.
150. KEEG Genes Database [Internet]. [cited 2019 Mar 9]. Available from: <https://www.genome.jp/kegg/genes.html>
151. Badui S. Química de los alimentos. 5ta ed. México: Pearson Educación de México; 2012. 744 p.
152. Sumano H, Ocampo L. Farmacología Veterinaria. 2da ed. México: McGraw - Hill; 1997. 1092 p.
153. Sumano H, Gutiérrez L. Farmacología clínica en aves comerciales. 4ta ed. México: McGraw - Hill; 2010. 713 p.
154. Castilla FA. Efecto de la inclusión de un suplemento nutricional líquido sobre los parámetros productivos según la edad de pollos de engorde. Universidad Nacional Mayor de San Marcos; 2018.
155. Memon M, Hussain I, Sahito H, Abbasi F, Iqbal M, Memon A. Effect of biotin as a feed additive on the growth of broiler. *Int J Innov Agric Biol Res*. 2014;2(4):17–25.
156. Lea HK. Gut health of poultry in the post-antibiotic era: elucidating the mechanisms underlying successful feed supplements. Nottingham Trent University; 2013.
157. Psifidi A, Fife M, Howell J, Matika O, van Diemen PM, Kuo R, et al. The genomic architecture of resistance to *Campylobacter jejuni* intestinal colonisation in chickens. *BMC Genomics* [Internet]. 2016 Apr 18;17:293.
158. He H, Genovese KJ, Kogut MH, Swaggerty CL, Pevzner IY. Selection for pro-inflammatory mediators produces chickens more resistant to *Campylobacter jejuni*. *Poult Sci* [Internet]. 2017 Feb 22;96(6):1623–7.
159. Li JH, Liu X, Xing L, Liu H, Li X, Bao J. Gene Expression Profiling of Broiler

Liver under Cold Stress by High-Throughput Sequencing Technology. Vol. 54, The Journal of Poultry Science. 2017.

160. Kang SW, Kuenzel WJ. Regulation of gene expression of vasotocin and corticotropin-releasing hormone receptors in the avian anterior pituitary by corticosterone. *Gen Comp Endocrinol* [Internet]. 2014;204:25–32.
161. Gu X, Feng C, Ma L, Song C, Wang Y, Da Y, et al. Genome-Wide Association Study of Body Weight in Chicken F2 Resource Population. *PLoS One* [Internet]. 2011 Jul 14;6(7):e21872.
162. Kong B-W, Hudson N, Seo D, Lee S, Khatri B, Lassiter K, et al. RNA sequencing for global gene expression associated with muscle growth in a single male modern broiler line compared to a foundational Barred Plymouth Rock chicken line. *BMC Genomics* [Internet]. 2017 Jan 13;18(1):82.
163. Huang QC, Xu ZR, Han XY, Li WF. Effect of dietary betaine supplementation on lipogenic enzyme activities and fatty acid synthase mRNA expression in finishing pigs. *Anim Feed Sci Technol* [Internet]. 2008;140(3):365–75.
164. How the RNA Integrity Number (RIN) Works [Internet]. 2004. Available from: http://naes.unr.edu/genomics/documents/Agilent-How_RNA_Integrity_Number_Works.pdf
165. Advantages of paired-end and single-read sequencing [Internet]. [cited 2019 Jun 6]. Available from: <https://www.illumina.com/science/technology/next-generation-sequencing/paired-end-vs-single-read-sequencing.html>
166. Sengupta S, Bolin JM, Ruotti V, Nguyen BK, Thomson JA, Elwell AL, et al. Single read and paired end mRNA-Seq Illumina libraries from 10 nanograms total RNA. *J Vis Exp* [Internet]. 2011 Oct 27;(56):e3340–e3340.
167. Trivedi UH, Cézard T, Bridgett S, Montazam A, Nichols J, Blaxter M, et al. Quality control of next-generation sequencing data without a reference. *Front Genet* [Internet]. 2014 May 6;5:111.
168. Chicken assembly and gene annotation [Internet]. 2016 [cited 1BC Mar 6].

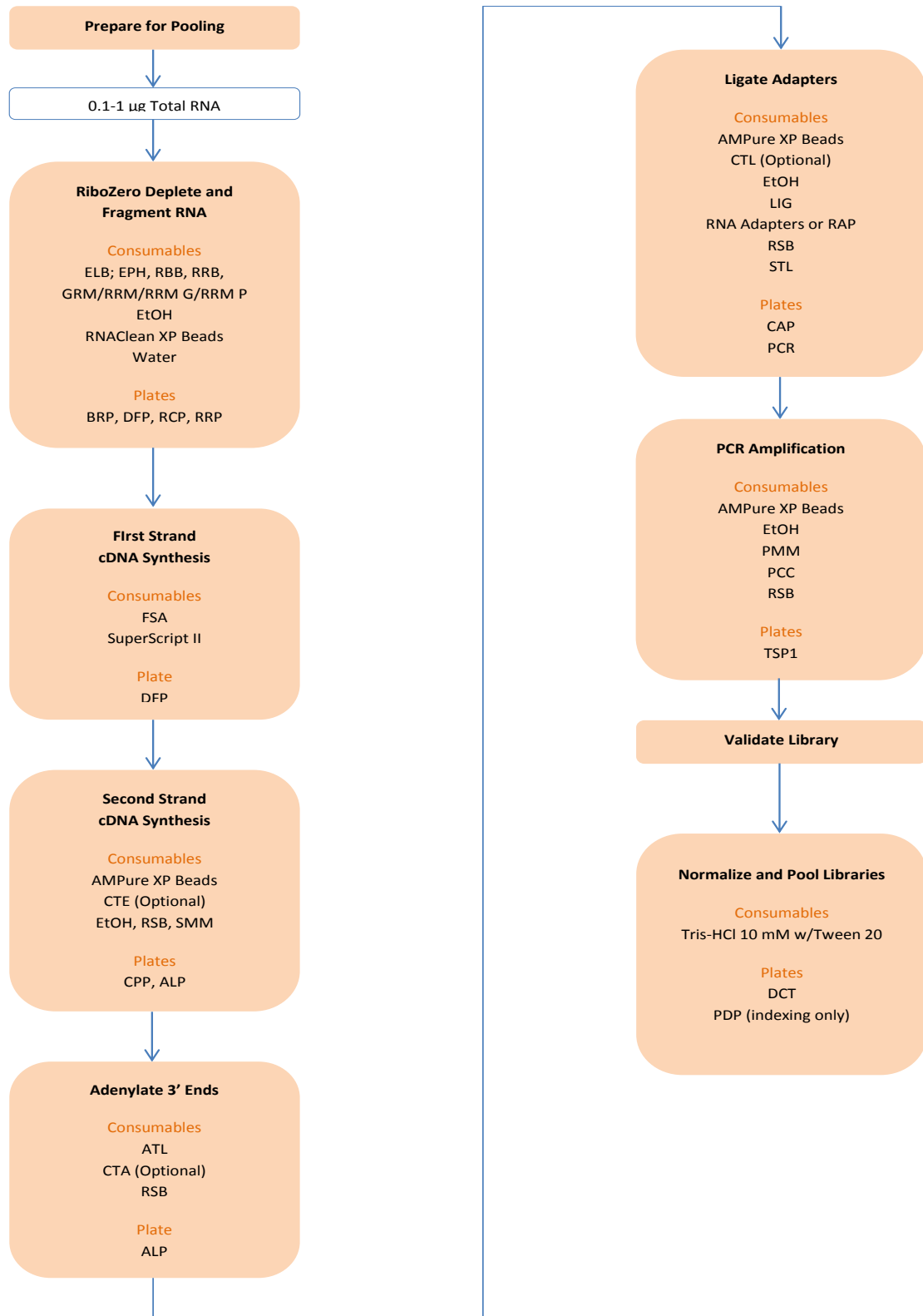
169. Yi B, Chen L, Sa R, Zhong R, Xing H, Zhang H. Transcriptome Profile Analysis of Breast Muscle Tissues from High or Low Levels of Atmospheric Ammonia Exposed Broilers (*Gallus gallus*). PLoS One [Internet]. 2016 Sep 9;11(9):e0162631.

IX. ANEXOS

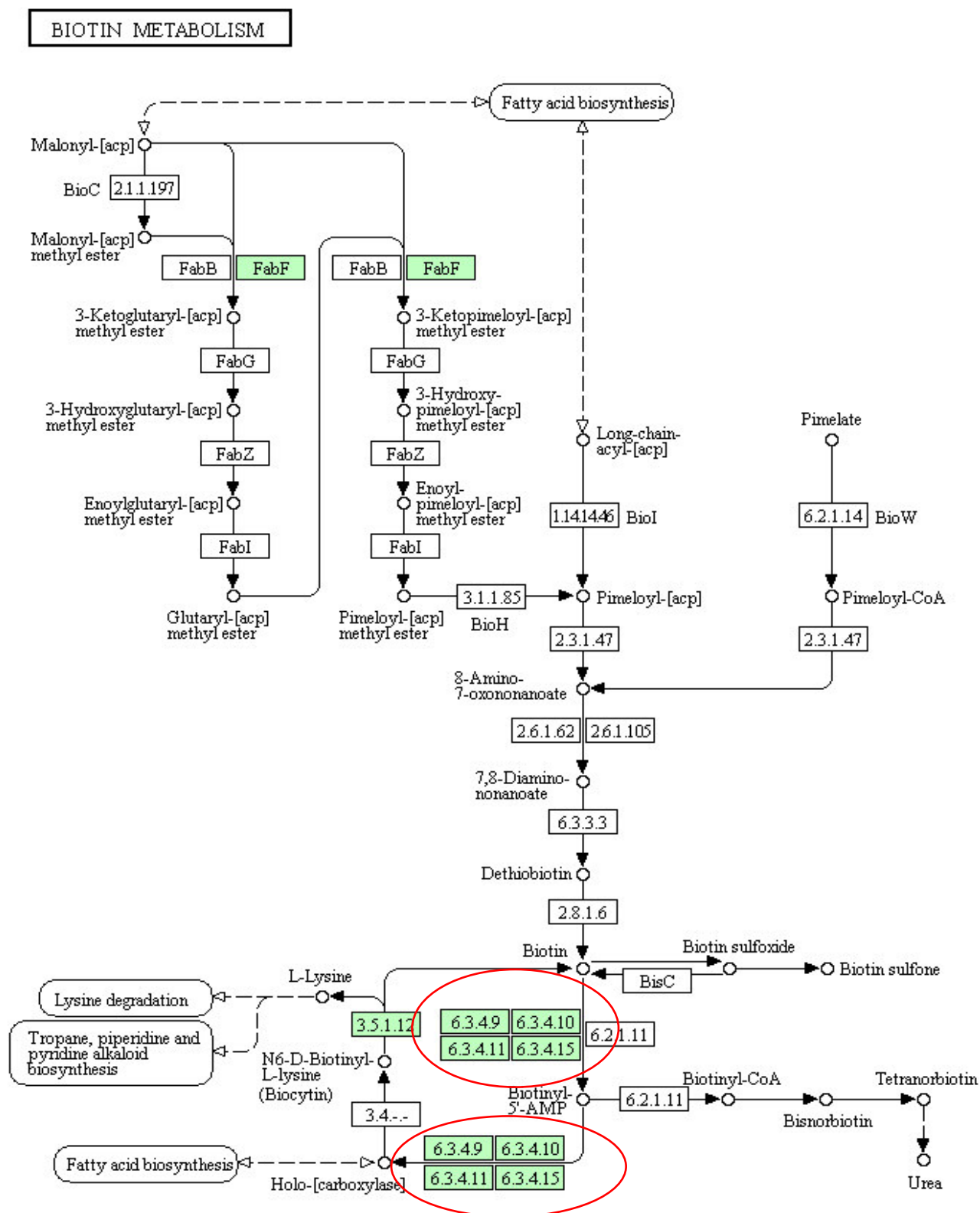
Anexo N°1. PROCEDIMIENTO DE EXTRACCIÓN DE ARN DE TEJIDO ADIPOSO DE *Gallus gallus*

1. Pesar 2 muestras de tejido adiposo extraídas desde la pechuga. El peso de tejido osciló entre 150 a 300 mg.
2. Homogenizar la muestra lo mayor posible para que el tejido se mantenga en porciones lo más pequeñas posibles.
3. Agregar el tejido a un tubo eppendorf de 1.5ml y junto con este 1ml del reactivo TRIzol. Homogenizar.
4. Agregar 200µl de cloroformo.
5. Llevar los tubos con las muestras al vórtex. Aproximadamente de 2 a 3 minutos.
6. Llevar las muestras a -20°C durante 15 minutos.
7. Centrifugar las muestras a 11000 rpm por 18 minutos y temperatura ambiente.
8. La fase acuosa que se genera luego de la centrifugación se traslada a otro tubo. Esta fase debe tener un color rosado bastante claro, casi transparente.
9. Agregar 500 µl de isopropanol.
10. Llevar las muestras a -20°C durante 15 minutos.
11. Centrifugar las muestras a 11000 rpm por 20 minutos y a 4°C.
12. Remover el sobrenadante.
13. Lavar el pellet con etanol al 75%.
14. Centrifugar a 10000 rpm por 5 minutos a 4°C.
15. Dejar secar por 30 minutos aproximadamente.
16. Agregar 30µl de agua libre de RNasas y DNasas.
17. Verificar la correcta extracción mediante electroforesis y vista al transiluminador LED de luz azul.

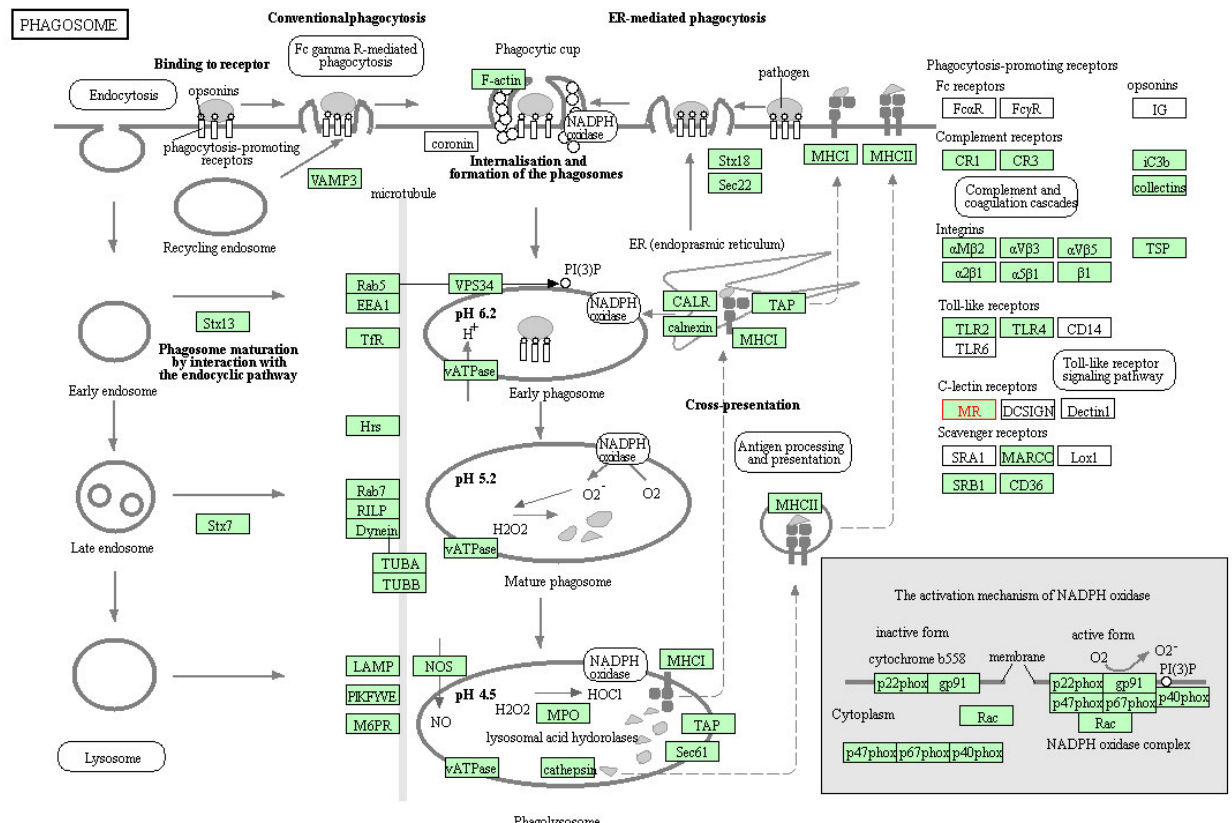
Anexo N°2. Preparación de la muestra de ARN total por el protocolo TruSeq (flujo de trabajo para una cantidad de muestras menor a 24). Tomado de la Guía para preparación de la muestra: TruSeq Stranded ARN Total.



Anexo N°3. Metabolismo de la Biotina, donde participa la enzima holo-carboxilasa (HLCS) codificada por el gen del mismo nombre que se encuentra sobre-regulado en la muestra CA001 con respecto a la muestra MB001 (p ajustado < 0.05). Se indica su ubicación en el metabolismo dentro de los círculos rojos. Fuente: KEGG



Anexo N°4. Ruta de formación del fagosoma, donde participa el gen MARC1 (receptor de manosa C, tipo 1), un gen sub-regulado en la muestra CA001 con respecto a la muestra MB001 (p ajustado < 0.05). Lo podemos ver resaltado de rojo.



Fuente: KEGG